EVALUATING THE PERFORMANCE OF EXISTING AND NOVEL EQUIVALENCE TESTS FOR STRUCTURAL EQUATION MODELING

NATALY BERIBISKY

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ON

AUGUST 2023

ABSTRACT

It has been suggested that equivalence testing (otherwise known as negligible effect testing) be used to evaluate model fit within structural equation modeling (SEM). This dissertation is composed of two studies that propose novel equivalence tests based on the popular RMSEA, CFI and SRMR fit indices. Using Monte Carlo simulations, each study compares the performance of these novel tests to other existing equivalence testing-based fit indices in SEM, as well as to other methods commonly used to evaluate model fit. In each study, results indicate that equivalence tests in SEM have good Type I error control and display considerable power for detecting well-fitting models in medium to large sample sizes. At small sample sizes, relative to traditional fit indices, equivalence tests limit the chance of supporting a poorly fitting model. Both studies also present illustrative examples to demonstrate how equivalence tests can be incorporated in model fit reporting. We recommend that equivalence tests be utilized in conjunction with descriptive fit indices to provide more evidence when evaluating model fit.

ACKNOWLEDGMENTS

Graduate school was beyond anything I could have hoped for because of my supervisor Rob Cribbie. Rob has been an incredible mentor throughout my graduate studies, always providing ample knowledge, kindness, patience, and support. Rob created an environment where I felt supported in trying to achieve any milestone, and where I felt that I was able to grow as a researcher. I am incredibly grateful for always being able to rely on him for any kind of assistance. Rob, it is impossible to thank you enough for challenging me, believing in me, and making me the researcher I am today.

I am also immensely grateful to Dave Flora. Taking Dave's course in structural equation modeling was what first inspired my interest in the topic, and ultimately led to structural equation modeling being the focus of my dissertation research. Dave has been an incredible teacher during my graduate studies, both when I was a student in his courses and when he provided expert feedback on all stages of my dissertation. I am so appreciative for the insight that Dave's feedback has brought me and am grateful to have the opportunity to learn from him.

Next, I am thankful for Monique Herbert, whose comments at all stages of the dissertation always provided me with clarity. Monique consistently challenged me to think about the practical implications of my project and consider the connections between my dissertation and statistical pedagogy. I am grateful to have learned from Monique's approach to understanding statistics and thinking about topics in a "big-picture" context.

I would also like to thank my examining committee. Thank you to Cathy Zhang for chairing the committee. I was also lucky to have Cathy on my committee as some of her past research was directly relevant to my project. I am grateful to my external examiner Andrea Howard for challenging me to think about how my work connected with the broader field of

statistics and how researchers will use the tools created by quantitative methodologists. Finally, thank you to Georges Monette for his thoughtful contributions to the dissertation discussion which made me view my project in a different light. I am also immensely grateful to have had the pleasure of working with Georges in the Statistical Consulting Service at York. I have learned so much from being able to listen to him deconstruct various consulting cases.

Thank you to Phil Chalmers (and SimDesign!). I am so lucky to have been able to talk through the particulars of the dissertation's simulations with Phil. I have greatly benefited from Phil's kindness, whether it was rerunning simulations or just talking through what was going on behind the scenes in the computing process. I am also grateful to Phil for coming to my defence.

I would also like to thank Gregory Hancock. It has been surreal for me to get the opportunity to work with Greg. It has been such a pleasure to get to learn from Greg's approach to research. I am grateful to Greg for his continued kindness and support and am also appreciative that he was able to virtually attend my defence.

Thank you so much to the Cribbie Lab! Thanks to Linda; I am so lucky that we started the program together and got to share so many experiences (first conference talks, first research projects, etc.). Thank you to Udi for the opportunity to start coding club together and always providing me with support. Thanks to Naomi for being incredibly supportive, caring, helpful, and fun to talk to. Though not in the Cribbie Lab, thanks to Stephan for all your encouragement.

I am extremely indebted to my friends who have provided boundless joy throughout my graduate school experience. Emma, thank you for always working/studying with me (both virtually and person!) and being an incredible source of support, whether it is on a recipe I am making or a paper I am submitting. It is always such a joy to spend time with both you and Jamie; thank you both for always being there. Branislav and Maša, you have both been such a

source of stability during my degree. Thank you both for always believing in me and consistently being there for me. Branislav, it was such a wonderful surprise to have you at my defence, but upon reflection, it really could not have been any other way. Thank you to Gili and Lidiya, your support and excitement has always been felt from near and far. Thank you to Mahsa and Nabil, Mahsa, it has been such an incredible experience to go through our academic journeys side by side. Thank you to Dana and Pritish, Dana, your encouragement has always brightened my day.

I am incredibly grateful to my in-laws, Avi and Marlene for their endless encouragement and support. Thank you both for filling every moment with laughter, for being encouraging every step of the way, and for your love. I am so lucky to be your daughter-in-law. To Alex, my brother-in-law, thank you for flying in for my defence, and thank you for always being someone I can count on, whether it be a big life event or sharing a seat on Big Thunder Mountain. Thank you to my brother Alex and my sister-in-law Masha. Masha, thank you for all your support and encouragement throughout my degree. Alex, thank you for your support, as well as paving the way and being such an example to look up to. To my mom, you have been a source of strength and inspiration. Your support has been the home I have been always lucky to return to. Thank you for being excited with me during the big moments and the little ones.

To Nathan, you are my anchor and my home. This would not have been possible without you, from your help in understanding mathematical concepts to your endless patience with me. You are the best partner one could ever ask for.

Finally, I dedicate this thesis to my late father, who, when I was little and asked for help with my homework, would reply by always asking whether I really understood the question. Doing research has made me understand just how important your response was. I miss you and hope to make you proud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

GENERAL INTRODUCTION

Structural equation modeling (SEM) allows researchers to create multivariate models that estimate how different variables (both observed and latent) are related to one another. Before these associations are interpreted, the fit of the model to the data is evaluated. Evaluating model fit is a multi-faceted process that usually involves the inspection of various fit indices. Fit indices provide different descriptive evaluations of a model's fit to data. The three most common fit indices reported (see Jackson et al., 2009) include the root mean squared error of approximation (RMSEA: Steiger & Lind, 1980), the comparative fit index (CFI: Bentler, 1990), and the standardized root mean squared residual (SRMR: Bentler, 1995).

All fit indices provide various descriptive information about a model's fit to data. For instance, this information could estimate the true amount of misfit in the population (i.e., RMSEA), compare the fit of a researcher's model to one without any associations among variables (i.e., CFI), or measure the average value of residual correlation left behind once the model has been fit to the data (i.e., SRMR). Because they are descriptive values, fit indices were intended to be interpreted as effect sizes, allowing the researcher to evaluate how well their model fit the data, with each fit index being interpreted along its own unique range. In practice, researchers often compare fit indices with various cut-off values. These cut-off values are perceived as the boundary between fit being acceptable or not. Thus, in practice, instead of being interpreted as effect sizes, fit indices are sometimes quickly compared to a cut-off value, like a pseudo-hypothesis test, without the value itself being meaningfully interpreted. This process clouds the intended use of fit indices. We term the procedure of comparing a fit index to a cut-off an informal check test (ICT).

Prior to recent developments in the SEM literature, the only inferential test available to researchers when evaluating model fit was the χ^2 goodness of fit test. Unfortunately, this test has been greatly critiqued for having an unrealistic null hypothesis, with many researchers recommending against its use (e.g., Bentler & Bonett, 1980; Browne & Cudeck, 1992; Steiger, 2007). Thus, there has existed a need for inferential tests that are distinct from the χ^2 goodness of fit test and may disentangle fit indices from being used only as ICTs.

One such approach to inferential testing in SEM is equivalence testing. Equivalence testing has been incorporated into model fit evaluation with tests for RMSEA (MacCallum et al., 1996) and CFI (Yuan et al., 2016). All equivalence tests for fit indices compare the amount of misspecification in a given model to a predetermined tolerable size of misspecification. In practice, this involves comparing a given bound of a confidence interval (CI) around a fit index to an equivalence bound or EB (one bound of an equivalence interval).

Although equivalence tests for CFI have been proposed (Yuan et al., 2016), their performance has not yet been evaluated alone or in tandem with RMSEA. Further, despite SRMR being one of the most reported fit indices (Jackson et al., 2009), it does not yet have a corresponding equivalence test. To address these opportunities for additional research, the following dissertation is composed of two studies. The first study, described in Chapter 2, focuses on RMSEA and CFI and compares the performance of extant equivalence tests (i.e., MacCallum et al., 1996; Yuan et al., 2016) to newly proposed equivalence testing procedures through a Monte Carlo simulation study. The second study, which can be found in Chapter 3, proposes new equivalence tests for SRMR and compares all proposed variations of the tests using a Monte Carlo simulation study. In each study, common methods for evaluating model fit are also included in the simulation (namely, ICTs and the χ^2 goodness of fit test). Each study also

includes an illustrative example which applies the equivalence tests to an SEM model with real data.

CHAPTER TWO

EQUIVALENCE TESTING BASED FIT INDEX: ROOT MEAN SQUARED ERROR OF APPROXIMATION AND COMPARATIVE FIT INDEX (STUDY 1)

Structural equation modeling (SEM) allows researchers to specify and estimate how observed variables are associated with each other and latent constructs. Specifically, researchers can hypothesize a model that specifies how the observed variables and hypothesized constructs are related. To evaluate whether a proposed model is a reasonable depiction of the structural associations underlying observed data, the model can be examined with various goodness of fit indices. Fit indices, such as the root mean squared error of approximation (RMSEA) or comparative fit index (CFI) allow a researcher to judge how well a model's proposed structure represents the data. These fit indices may assess model fit descriptively in different ways: For instance, they could do so by evaluating the amount of discrepancy error per the specified model's degree of freedom (RMSEA: Browne & Cudeck, 1992) or by comparing the fit of the observed model to a null model with no associations among the variables (CFI: Bentler, 1990).

Reporting a set of fit indices is meant to give the reader information about different aspects of model fit. Recent work has been conducted on creating effect size measures that capture varying degrees of model misspecification (see, for instance, Gomer et al., 2019). Yet, in their conception, fit indices themselves were meant to function as effect sizes, offering nuanced interpretations of model fit along a continuum (McNeish & Wolf, 2021). Unfortunately, descriptive fit indices are often used solely as informal check tests (ICTs), where their magnitude is compared to certain cut-off values to distinguish whether the fit of a model is satisfactory. In this vein, instead of being used as effect sizes, fit indices are at risk of becoming purely quasi-significance tests, with model fit simply being viewed as acceptable or not.

Traditionally, the only null hypothesis significance test that accompanies an SEM model is a chi-square (χ^2) goodness of fit test. The χ^2 goodness of fit test evaluates whether the data's observed covariance matrix matches the model-implied covariance matrix (Bollen, 1989). In order to support a hypothesized model, the aim of this test is to find a non-significant test statistic, since the null hypothesis is that the model's structure perfectly represents the structure of the data (Marcoulides & Yuan, 2017). However, a consequence of using the χ^2 goodness of fit test, which is highly sensitive to sample size, is that even if one obtains a non-significant test statistic, they can only *fail to reject* the null hypothesis that the two covariance matrices are identical (Yuan et al., 2016). In other words, the χ^2 goodness of fit test evaluates an unrealistic null hypothesis (because no model is perfect) and aims to accept this null hypothesis (which in itself is a problematic strategy, as we will illustrate in detail). The use of descriptive fit indices, such as RMSEA and CFI, is meant to circumvent the problems of the χ^2 goodness of fit test. However, instead of functioning as effect sizes, these fit indices themselves often become ICTs by being directly compared to cut-off values in an attempt to make an inference regarding model fit (e.g., Browne & Cudeck, 1992; Hu & Bentler, 1999; MacCallum et al., 1996).

While significance tests for model fit are secondary to interpreting the magnitude of the descriptive fit indices themselves, they offer additional information about goodness of fit that may be useful to researchers. However, the two mentioned approaches (the χ^2 goodness of fit test and ICTs) are both problematic procedures for assessing model fit in SEM (we provide details later). Accordingly, to evaluate whether the fit of a model is satisfactory with a formal significance testing framework, it is necessary to look beyond these approaches.

Equivalence testing, or negligible effect testing, provides researchers with the framework to detect a lack of association among variables (Counsell & Cribbie, 2015), a lack of difference

in means (Rogers et al., 1993; Wellek, 2010), and recently has been proposed for evaluating a model's fit to data (Yuan et al., 2016). These equivalence tests have been proposed for RMSEA (MacCallum et al., 1996) and CFI (Yuan et al., 2016) and can be used to supplement the reporting of different descriptive fit indices and potentially replace the χ^2 goodness of fit test.

The use of equivalence testing in evaluating a model's hypothesized structure gives researchers many further insights and advantages beyond those available with a traditional null hypothesis testing framework via the χ^2 goodness of fit test (Marcoulides & Yuan, 2017) or using ICTs. The proposed study adds to the literature on evaluating model fit in SEM via equivalence testing in three ways. First, we present variations to the two extant equivalence tests for RMSEA and CFI. Second, we use SEM models common in the social and behavioural science literature to evaluate the performance of the novel and extant equivalence tests, as well as compare the tests' performance to ICTs and the χ^2 goodness of fit test. Lastly, we present an example of how equivalence tests can be implemented in the open-source software R (R Core Team, 2021). The example uses real data to illustrate how equivalence tests for CFI and RMSEA can be interpreted and reported. We also describe how to interpret these tests using confidence intervals (CIs).

Traditional Fit Indices

After a researcher specifies a model's structure, the model is estimated such that the observed covariance matrix from sample data is compared to the researcher-specified (or model-implied) covariance matrix based on some criterion. With maximum likelihood (ML), estimation is accomplished using the discrepancy function F_{ML}:

$$F_{ML} = \ln\left|\widehat{\boldsymbol{\Sigma}}\right| - \ln|\mathbf{S}| + \mathrm{Tr}\left(\mathbf{S}\,\widehat{\boldsymbol{\Sigma}}^{-1}\right) - p$$

where $\widehat{\boldsymbol{\Sigma}}$ is the model-implied covariance matrix, $\mathbf{S}$ is the observed covariance matrix, and p is the number of manifest (observed) variables (note, | | refers to the determinant of the matrix, Tr()

refers to the trace of the matrix, and matrices and vectors are denoted in bold). When $\mathbf{S}$ and $\widehat{\boldsymbol{\Sigma}}$ are identical, $F_{ML} = 0$.

This estimator is based on an assumption that the vector $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This multivariate normality assumption further allows the use of a likelihood-ratio χ^2 statistic to evaluate the null hypothesis that the observed covariance matrix is equal to the model-implied covariance matrix, giving the χ^2 goodness of fit test discussed above. Specifically, if the model is correctly specified, χ^2 asymptotically follows a χ^2 distribution with degrees of freedom $df = \frac{p(p+1)}{2} - q$, where q is the number of distinct parameters requiring estimation. The traditional χ^2 goodness of fit test has the following null hypothesis:

$$H_0: \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}},$$

(Browne & Cudeck, 1992). Thus, the null states that the population and model-implied covariance matrices are equal. Consequently, when the null hypothesis is false by any degree, having a reasonable sample size should result in rejecting the null hypothesis (Bentler & Bonett, 1980). In contrast, in small samples, the χ^2 goodness of fit test may not have enough power to reject a false null hypothesis (i.e., distinguish between models that fit well and those that do not, Kenny & McCoach, 2003). In any case, the null hypothesis of this test is largely regarded as unrealistic (e.g., Browne & Cudeck, 1992; Steiger, 2007; Hooper et al., 2008). Accordingly, instead of the χ^2 goodness of fit test, merit is given to the interpretation of fit indices such as the RMSEA and CFI (which utilize the likelihood ratio χ^2 statistic in their calculation).

RMSEA

The RMSEA index was first defined by Steiger and Lind (1980). This fit index captures errors of approximation (Browne & Cudeck, 1993; Steiger & Lind, 1980), which can be defined

as the amount of misfit between the model-implied covariance matrix and the population covariance matrix. More specifically, RMSEA incorporates weighted sums of squared deviations between these matrices (Rigdon, 1996).

Because the population covariance matrix is unknown in practice, a sample RMSEA statistic is approximated using the sample covariance matrix. RMSEA penalizes models that are more complex by including degrees of freedom in its calculation. Models with fewer parameters to estimate (higher degrees of freedom) will lead to lower values of RMSEA, which indicates a better fitting model. RMSEA has a range from zero to positive infinity. Population RMSEA can be defined as:

$$\text{RMSEA}_{\text{pop}} = \sqrt{\frac{F_{\text{ML}}}{df}}.$$

F_{ML} may also be defined in terms of the non-centrality parameter, λ, of a non-central χ^2 distribution, which quantifies population model misfit (Nevitt & Hancock, 2000). Specifically, λ is the product of $F_{\text{ML}}(N - 1)$. Thus, the sample RMSEA can be calculated as:

$$\text{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df\,(N - 1)}},$$

where $\hat{\lambda}$ is an estimate of the non-centrality parameter λ (note that $\hat{\lambda} = \chi^2 - df$, where χ^2 represents the likelihood ratio χ^2 statistic). Thus, sample RMSEA estimates true model misfit within the population (Nevitt & Hancock, 2000). Under model misspecification, $F_{\text{ML}}(N - 1)$, the likelihood ratio χ^2 goodness of fit statistic, follows a non-central χ^2 distribution, under the assumption that population model misfit is of approximately the same magnitude as sampling error (see MacCallum et al., 1996). With the caveat that RMSEA = 0 when the likelihood ratio χ^2 statistic is smaller than the degrees of freedom, asymptotically the RMSEA's distribution can be conceptualized as the square root of the rescaled non-central χ^2 distribution (that may be

identified with given degrees of freedom, sample size, and λ; Chen et al., 2008, Curran et al., 2003). However, note that under certain conditions (e.g., small to moderate sample sizes, non-normality), the likelihood ratio χ^2 statistic may not be central or non-central chi-square distributed (Yuan et al., 2015).

The association between the non-central χ^2 distribution and the RMSEA allows for the construction of CIs (MacCallum et al., 1996), which become important in equivalence testing. Specifically, once a model's RMSEA index is estimated, the RMSEA estimate and model degrees of freedom can be used in a nonlinear equation to iteratively solve for the upper and lower limits of the corresponding non-central χ^2 distribution (Browne & Cudeck, 1993). The upper and lower limits of the non-central χ^2 distribution ($\hat{\lambda}_L$ and $\hat{\lambda}_U$) can then be used to estimate the upper and lower limits of the CI (Browne & Cudeck, 1993):

$$CI_{lower} = \sqrt{\frac{\hat{\lambda}_L}{df\,(N-1)}}$$

$$CI_{upper} = \sqrt{\frac{\hat{\lambda}_U}{df\,(N-1)}}$$

where $\hat{\lambda}_L$ and $\hat{\lambda}_U$ represent the lower and upper limits of the non-centrality parameter for a given degrees of freedom value.

CFI

The CFI is defined such that a target or hypothesized model is compared to a null model for the same number of observed variables. In this null model, the covariance between any two variables is normally set to zero (with some exceptions in software for certain models). The CFI's range is usually between 0 and 1, but larger values are possible. The population CFI can be defined as:

$$CFI_{pop} = 1 - \frac{\lambda_t}{\lambda_i},$$

where λ_t is the noncentrality parameter corresponding to the hypothesized model and λ_i is the noncentrality parameter corresponding to the null model.

The formula for the sample CFI is:

$$\text{CFI} = 1 - \frac{\hat{\lambda}_t}{\hat{\lambda}_i},$$

where $\hat{\lambda}_t = \max(\chi^2_t - df_t, 0)$ and $\hat{\lambda}_i = \max(\chi^2_i - df_i, 0)$ where χ^2_t is the likelihood ratio statistic for the hypothesized model, df_t represents the associated model degrees of freedom, χ^2_i is the null model likelihood ratio statistic, and df_i represents the degrees of freedom for the null model (Bentler, 1990).

This formula illustrates that the smaller the likelihood ratio χ^2 statistic of the observed model (relative to that of the baseline model), the closer the CFI will be to one. Accordingly, higher CFI values are indicative of better fit. In comparison to the RMSEA fit index, CFI has a much more complicated association with the non-central χ^2 distribution. Specifically, Browne and Cudeck (1992) note the complexity of comparative fit indices that involve a comparison with a null model because, "even approximations to their distributions are seldom available and their values depend on the particular null model chosen" (p. 231). Accordingly, the calculation of a CI for CFI is more challenging, given that it involves both the hypothesized and null models.

Yuan et al. (2016) introduced a strategy for calculating the lower bound of the CI of the CFI by estimating the two non-centrality parameters specified above: one for the null model and one for the hypothesized model. First, the null model likelihood ratio χ^2 statistic and associated degrees of freedom are used to estimate λ for the null model. This is repeated for the hypothesized model, where the hypothesized likelihood ratio χ^2 statistic and the associated degrees of freedom are used to estimate the hypothesized λ (the exact method for obtaining λ applies a Cornish-Fisher inverse expansion to a set of quantities called cumulants, which are an

alternative to the moments of a probability distribution; this method is described in detail in Venables, 1975). Once the non-centrality parameter for each model (null and hypothesized) is estimated, CIs for the CFI can be estimated using a formula that Yuan et al. derive using set theory. Accordingly, Yuan et al.'s formula for the lower-bound of the CI for the CFI (for a CFI estimate bounded between 0 and 1) is:

$$CI_{lower} = 1 - \frac{\hat{\lambda}_t}{\max(\hat{\lambda}_t,\ \hat{\lambda}_i)}.$$

Equivalence Testing

Equivalence testing was originally made popular in the biopharmaceutical discipline, where researchers were trying to determine whether the difference in the effects of two drugs was so small that it was negligible or tolerable (i.e., drug type is minimally associated with effectiveness). Equivalence testing was presented to the social and behavioural sciences through instrumental articles such as Rogers et al. (1993) and Tryon (2001). Equivalence testing is preferable to traditional null hypothesis significance testing when a researcher aims to detect a lack of association among variables.

For instance, recall that the null hypothesis of the χ^2 goodness of fit test represents that of a perfectly fitting model:

$$H_0: \boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}},$$

Increasing the sample size will increase the probability of this null hypothesis being rejected, even at small levels of misspecification. This property is problematic in the context of SEM because no model will perfectly represent the features of a dataset; therefore, some misspecification is inevitable and can even be valuable for stimulating future research (Cudeck & Henly, 1991; MacCallum & Tucker, 1991).

When the null hypothesis of the χ^2 goodness of fit test is not rejected, this failure to reject the null is not evidence that the null is correct (Rogers et al., 1993). Therefore, the use of the χ^2 goodness of fit test can never directly demonstrate that the model-implied and observed covariance matrices are equal.

Equivalence Testing in Structural Equation Modeling

When incorporated into SEM for determining model fit, equivalence tests compare a model's misspecification to a minimally tolerable size of misspecification. It is possible to work in the scale of the discrepancy function (F_{ML}) or in the scale of the fit index of interest (e.g., RMSEA, CFI). Yuan et al. (2016) discuss using the tolerable size of misspecification for the discrepancy function (*T*-size) and transform the *T*-size to the scale of RMSEA and CFI. We chose to work in the scale of the fit indices because most applied researchers are familiar with values in the scale of RMSEA or CFI.

The tolerable size of misspecification may be approximated by a given value (note that the minimally tolerable size of misspecification is analogous to the minimally meaningful effect size, or MMES, in the equivalence testing literature). This value can be conceptualized as one bound of an equivalence interval, or an equivalence bound (EB). Any value exceeding the upper value of the EB (EB_{upper}) for RMSEA or below the lower value of the EB (EB_{lower}) for CFI is too large of a misspecification to be tolerable. Accordingly, the null hypothesis for an equivalence-based test of model fit specifies that the degree of model misspecification is greater than some tolerable level:

$$H_0: RMSEA_{pop} \geq EB_{upper}.$$

This null hypothesis is rejected when a model's level of misspecification is tolerable, hence providing support for the alternate hypothesis:

$$H_1: RMSEA_{pop} < EB_{upper}.$$

Tests that have been already proposed in the field of equivalence testing for fit indices, as with other equivalence tests, compare the bounds of CIs to given EB values. These tests are expected to follow the equivalence testing logic where the symmetric 100(1- 2α)% CI (α is the nominal Type I error rate) is compared to the equivalence interval. This comparison corresponds with each bound of the equivalence interval retaining a Type I error rate of α (which is applicable for equivalence tests for fit indices, where only one bound of the CI and one bound of the equivalence interval are relevant).

Existing Equivalence Tests for Fit Indices

MacCallum et al. (1996) proposed the not-close fit test for RMSEA, which we term the EBF_{RMSEA} (or equivalence-based fit test for RMSEA). The EBF_{RMSEA} compares the upper bound of the CI for RMSEA (CI_{upper}) to EB_{upper} (i.e., the upper bound of an equivalence interval). If the upper limit of the CI is smaller than EB_{upper}, the null hypothesis of not-close fit (H_0: $RMSEA_{pop} \geq EB_{upper}$) can be rejected, thus providing support for a close fit of the specified model (H_1: $RMSEA_{pop} < EB_{upper}$) (MacCallum et al., 1996).

Yuan et al. (2016) introduced an equivalence test for the CFI, which we term $EBF\alpha_{CFI\text{-}A}$ (or the equivalence-based fit test for CFI with modified CIs using adjusted EBs; the inclusion of α in the test name refers to a modified CI, while the -A subscript refers to an adjusted EB, both of which are discussed below). This test compares the lower bound of the CI for CFI (CI_{lower}) to a given lower bound of an equivalence interval, EB_{lower} (recall that higher values of CFI are indicative of better fit). In similar logic to MacCallum et al., we can reject the null hypothesis:

$$H_0: CFI_{pop} \leq EB_{lower},$$

when the lower bound of the CI is greater or equal to EB_{lower}, implying that the observed model is a substantial improvement over the null model. When the lower bound of the CI is less than EB_{lower}, the evidence instead supports the alternative hypothesis:

$$H_1: CFI_{pop} > EB_{lower}.$$

Proposed Modifications to Existing Equivalence Tests for Model Fit

Alpha Level. Instead of using the traditional symmetric 100(1- 2α)% CI for comparison with the EB, the $EBF\alpha_{CFI\text{-}A}$ uses a Bonferroni-type correction to obtain the non-centrality parameter for two models (i.e., divides the tail α-level by 2). Even though there are two models, there is only one null hypothesis test conducted; specifically, looking at whether the lower bound of the CI for the CFI is less than EB_{lower}. By splitting the tail α-level in half, this method is instead consistent with a 100(1 – α)% CI, overall (with a nominal Type I error rate of α/2, since rejections cannot occur in the upper tail). In the present study, we suggest that the $EBF\alpha_{CFI\text{-}A}$ be modified to be more consistent with equivalence testing logic by omitting this Bonferroni-type correction. Specifically, without the Bonferroni-type correction, the entire α level is appropriately contained in the lower tail when comparing the lower bound of the CI to the EB. This modification corresponds nicely to the 100(1- 2α)% CI used in the equivalence testing literature. It has been demonstrated that the equivalence test based on the symmetric 100(1-2α)% CI will always have a Type I error rate at α and that therefore tests using the symmetric 100(1-α)% CI are unnecessarily conservative (Wellek, 2010). We term this modified CFI test the EBF_{CFI}. We compared the performance of $EBF\alpha_{CFI\text{-}A}$ and the EBF_{CFI} in our simulation study.

It is also worth noting that Yuan et al. (2016) align the equal sign (=) to the alternate hypothesis, instead of the null hypothesis (e.g., $H_1: CFI_{pop} \geq EB_{lower}$ instead of $H_0: CFI_{pop} \leq EB_{lower}$). Here, we have chosen to align the equal sign with the null hypothesis for all

equivalence tests in this paper, to be consistent with past equivalence testing literature (e.g., Wellek, 2010).

Bootstrapped Confidence Intervals. Another variant of an equivalence test for both fit indices is to use one bound of a bootstrapped CI, instead of computing the CI using the procedures outlined above. A bootstrap procedure creates multiple samples from a given dataset by repeatedly sampling from the data with replacement (see Elfron & Tibshirani, 1998). In each instance of sampling, an estimate of interest is computed. All the estimates can be combined into a distribution and used to determine a CI.

In this study, the Yuan et al. (2007) bootstrap (which we, like others, refer to as the YHY bootstrap) was chosen as a potential method for obtaining the bootstrap CI. Like the logic behind the RMSEA fit index, bootstrap methods in SEM should address the error of estimation (Preacher & Merkle, 2012). The YHY bootstrap improves on the traditional non-parametric bootstrap's ability to capture the error of estimation (Preacher & Merkle, 2012). The YHY transforms the data in a given sample before the bootstrap process with the goal that the non-centrality parameters in both the bootstrap sample and the full sample data are equivalent (Zhang & Savalei, 2016). The transformation has already been applied in the construction of bootstrap CIs for SEM fit indices (see Cheng & Wu, 2017; Zhang & Savalei, 2016). Accordingly, the YHY bootstrap is a logical choice for a bootstrap-based equivalence test (for RMSEA and CFI) as it has been demonstrated to have good coverage rates (Zhang & Savalei, 2016).

When estimating the value of RMSEA or CFI (rather than the bounds of the CI) in a bootstrap setting, it follows that the average of the estimates returned from the bootstrap samples should be used to compute the index, in place of the sample formulae for RMSEA and CFI presented above (see, for example, Gomer et al., 2019; Yuan & Marshall, 2004). Accordingly, if

bootstrap equivalence tests for fit indices are used, they logically may be accompanied by the average of the bootstrap values for estimates of the fit indices themselves.

It is possible to compute the YHY bootstrap in the `lavaan` package (Rosseel, 2012). In order to accurately capture the natural variability in bootstrapped parameter values, both inadmissible and admissible solutions are retained in `lavaan` (Rosseel, 2022, personal communication). In this vein, both kinds of solutions are also kept for the proposed bootstrap test. We compare the performance of these proposed bootstrap CI tests, which we term the $EBFB_{RMSEA}$ and the $EBFB_{CFI}$, to extant equivalence tests in our simulation study.

Modified and Unmodified Equivalence Bounds. The one-sided EBs that are used for comparison to the corresponding bounds of the CIs have previously been taken from popular guidelines for model fit (e.g., Browne & Cudeck, 1992; MacCallum et al., 1996). The most common guidelines for the fit indices were obtained from a large-scale simulation study by Hu and Bentler (1999) that found the magnitudes of fit indices that were best suited to differentiate between models with and without misspecification. Yuan et al. (2016) proposed equivalence tests using modified EBs instead of the original Hu and Bentler values (termed "adjusted cut-offs"). Specifically, Yuan et al. calculated these modified bounds using best subsets regression. These bounds account for sample size and model degrees of freedom for each unique model. The modified EBs allow more models to be classified as having good fit (compared to equivalence tests which use traditional EBs such as .95 for CFI or .05 for RMSEA). However, due to inflated Type I error rates (e.g., Counsell et al., 2020), we are hesitant to recommend the modified EBs suggested by Yuan et al. We identify the modified EB tests with an -A in the subscript (namely: $EBF_{RMSEA\text{-}A}$, $EBFB_{RMSEA\text{-}A}$, $EBF\alpha_{CFI\text{-}A}$, $EBF_{CFI\text{-}A}$, $EBFB_{CFI\text{-}A}$). We compare the performance of these tests with their unmodified EB counterparts.

Present Study

The purpose of this study is to explore the performance of existing and novel equivalence testing procedures for assessing model fit, as well as to compare the performance of these equivalence testing procedures to traditional fit assessment techniques. Although Hancock and Freeman (2001) evaluated the performance of MacCallum et al.'s (1996) not-close fit test for RMSEA (EBF_{RMSEA}), their application was for sample-size planning and not for assessing specific kinds of model misfit. In Nevitt and Hancock's (2000) study, although there is a comparison of tests of the EBF_{RMSEA} to other CI tests for RMSEA (i.e., close-fit and exact-fit tests) under conditions of non-normality, the only model analyzed is a three-factor, nine-indicator CFA model. Chen et al. (2008) present more complicated models (which are the models used in this study); however, they do not directly compare the performance of the EBF_{RMSEA} with ICTs and the χ^2 goodness of fit test. Despite having tutorials created that demonstrate its use (e.g., Marcoulides & Yuan, 2017; McNeish & Wolf, 2021), the performance of the $EBF\alpha_{CFI\text{-}A}$ proposed by Yuan et al. (2016) has not yet been compared to equivalence tests using traditional EBs. This study also contributes to the literature by testing the performance of the proposed equivalence tests based on the YHY bootstrap ($EBFB_{RMSEA}$, $EBFB_{RMSEA\text{-}A}$, $EBFB_{CFI}$, $EBFB_{CFI\text{-}A}$); although the performance of the YHY bootstrap has been explored with relation to SEM models (Zhang & Savalei, 2016; Cheng & Wu, 2017), it has yet to be translated into the equivalence testing for model fit literature.

Monte Carlo Simulation Study

The Monte Carlo study compared the performance of the various equivalence tests to one another, to the χ^2 goodness of fit test, and to the ICTs. With the combination of modified and unmodified EBs, and different kinds of CIs used for the tests (e.g., bootstrap vs direct estimation,

$1 - \alpha$ vs $1 - \alpha/2$), there are ten unique equivalence tests that we have chosen to compare in our Monte Carlo simulation study (definitions, CIs, and equivalence interval information for all equivalence tests are in Table 1).

We simulated data using models based on Chen et al. (2008). These models were selected because they were representative of SEM models found within social sciences (Chen et al., 2008). We used these models to evaluate the tests' performance with varying degrees of model misspecification across a wide range of sample sizes.

It is important to note that the methods that we compared investigate different research hypotheses, some aligned with the null hypothesis and some aligned with the alternate hypothesis. Specifically, the equivalence-based CFI tests evaluate whether the population CFI value falls above EB_{lower} (e.g., .95); the equivalence-based RMSEA tests evaluate whether the population RMSEA value falls below EB_{upper} (e.g., .05); the χ^2 goodness of fit test evaluates whether the population covariance matrix equals the model-implied covariance matrix; and the ICTs test whether the fit index itself is smaller (RMSEA) or larger (CFI) than an informal cut-off. Although the last procedure is not a formalized inferential test like the other methods, researchers often treat the result of an ICT as a binary accept-reject approach to model fit; therefore, we include it here.

We used a combination of the `simsem` package (Pornprasertmanit et al., 2020) and `SimDesign` package (Chalmers & Adkins, 2020) in R for the Monte Carlo study. `SimDesign` manages the structure of the simulation experiment using a three-step workflow procedure (Chalmers & Adkins, 2020), while `simsem` enables data analysis in SEM models in a Monte Carlo framework (Pornprasertmanit et al., 2020). Start values for the models were the defaults in the `lavaan` package (Rosseel, 2012); namely, latent variable variances were set to 0.05, observed

and residual variances were set to half of their observed sample variance, regression parameter start values were computed using ordinary least squares, covariances were set to 0, and the FABIN3 instrumental variable method was used to compute the start values of the factor loadings.

Population Data-Generating Models

Our Monte Carlo study generated data consistent with the three models depicted in Figures 1, 2, and 3. These models were adapted from Chen et al. (2008). Chen et al. arrived at these models' specifications using their own experience and reviewing prototypical models found in key journals. In this study, we have slightly adapted the population factor loadings from Chen et al. to create the three degrees of misspecification that we outline below. Each model contained three latent variables. Model 1 had three to four indicators for each latent variable for a total of nine indicators. Model 2 had four to six indicators for every latent variable for a total of fifteen indicators. Like Model 1, Model 3 had three to four indicators per latent variable, but also contained four correlated exogenous predictors of the latent variables.

The loadings used for the models in our Monte Carlo study differed slightly from Chen et al. (2008) because we were interested in the pattern of misspecification for both RMSEA and CFI fit indices (Chen et al. investigated patterns of RMSEA only; however, values of RMSEA and CFI for a single fitted model often lead to discrepant qualitative assessments of fit; see Lai & Green, 2016). We selected these loadings in order to define misspecified models that met criteria for conditions of negligible or non-negligible misspecification outside of the EB (i.e., both CFI > .95 and RMSEA < .05 for the negligible condition and both CFI < .95 and RMSEA > .05 for the non-negligible condition outside of the EB; note that by incorporating forms of misspecification which we describe in detail below, e.g., cross-loading omission, the difference between

$RMSEA_{pop}$ and the associated EB_{upper} was not equivalent to the difference between CFI_{pop} and the associated EB_{lower}; see Lai & Green). The population fit indices for the misspecified models are in Figures 1, 2, and 3. Model 1 had primary factor loadings of 0.70 in the standardized metric. Additionally, indicators that had a non-zero association with an additional latent variable had cross-loadings of 0.32. Model 2 had primary standardized factor loadings of 0.70, as well as additional cross-loadings (where indicated) of 0.35. Model 3 had standardized factor loadings of 0.70 cross-loadings of 0.21.

Test Specifications

The unmodified equivalence interval bound for RMSEA was set to .05. This was also the value used for the ICT comparison for the RMSEA, as well as the cut-point for determining whether a condition is a power condition or a Type I error condition with respect to the equivalence-based procedures (i.e., RMSEA < .05: power, RMSEA ≥ .05: Type I error). Using the formula provided by Yuan et al. (2016), the unique adapted RMSEA EB analogous to .05 was computed for every simulation condition. Because the adapted EB calculation depends on both sample size and model degrees of freedom, the adapted EB was different for every condition of the Monte Carlo study for a given model. As introduced in the start of the paragraph, the unmodified equivalence interval bound for CFI was set to .95. This was also the value used for the ICT comparison for the CFI, as well as the cut-point for defining a power condition or Type I error condition with respect to the equivalence-based procedures (i.e., CFI > .95: power, CFI ≤ .95: Type I error). Like for RMSEA, the modified CFI EB was separately computed for each condition of the Monte Carlo study.

Factors Manipulated in the Monte Carlo Study

Sample size and degree of misspecification were the two primary factors manipulated in the study. Sample size conditions included N = 50, 75, 100, 200, 400, 800, 1000, and 5000, which were also used in Chen et al. (2008) and represent sample sizes that are commonly found in the social and behavioural sciences.

Four kinds of model specification were applied to the each of the three population data-generating models: (1) perfect fit, (2) negligible misspecification, (3) non-negligible misspecification beyond the bound of the equivalence interval, and (4) non-negligible misspecification at the bound of the equivalence interval. In the perfect fit condition, the correct model structure was fit onto sample data drawn from each of the three population models. The negligible misspecification condition had fitted models with the following misspecifications: for Model 1, the cross-loading between indicator 7 and latent variable 2 was incorrectly fixed to 0 (i.e., omitted); for Model 2, the cross-loading between indicator 11 and latent variable 2 was omitted; and for Model 3, two cross-loadings (one between indicator 7 and latent variable 2 and another between indicator 6 and latent variable 3) were omitted. The non-negligible misspecification condition beyond the bound of the equivalence interval had the following additional misspecifications (beyond those of the negligible misspecification): Model 1 omitted two cross-loadings (one between indicator 4 and latent variable 1 and another between indicator 6 and latent variable 3); Model 2 omitted two cross-loadings (one between indicator 6 and latent variable 1 and another between indicator 10 and latent variable 3); and Model 3 omitted a cross-loading (between indicator 4 and latent variable 1) and a regression coefficient between an observed exogenous variable and a latent variable (between exogenous variable 3 and latent variable 2).

The non-negligible misspecification at the bound of the equivalence interval was configured differently. In order to investigate each equivalence test's Type I error rate when the relevant bound of the CI was aligned with the EB, we simulated two scenarios for each of the three models (one for RMSEA and one for CFI). In these Type I error scenarios, we created conditions such that the underlying model fit corresponded as closely as possible to either an RMSEA index of .05 (scenario 1) or a CFI of .95 (scenario 2). To create these scenarios in Models 1 and 2, we used the original population data-generating models and altered the size of the cross-loadings. We then fit models from the non-negligible misspecification beyond the equivalence bound condition (already described above) to sample data drawn from the modified population-generating models (fit was improved from the non-negligible condition described in the previous paragraph because we decreased the size of the cross-loadings). For Model 3, we altered the fitted models in the negligible specification condition (rather than the population models) by fixing one of the coefficient values to a different magnitude than in the population model, instead of allowing it to be freely estimated. In each scenario, we confirmed that the RMSEA or CFI values were at the EB by fitting the models to the corresponding population covariance matrices; analogous RMSEA or CFI values were exactly .050 or .950, respectively, to three decimal places.

Power and Error Conditions Across Misspecifications

Below, we describe how the misspecification conditions in our Monte Carlo study relate to power and error rates in equivalence testing, the χ^2 goodness of fit test, and ICTs. For a detailed discussion of differences between power and error rates in equivalence testing and traditional null hypothesis significance tests, see Yuan et al. (2016). See Figure 4 for the

locations of the population values of the fit indices relative to the EBs across the misspecification conditions.

In the perfect fit condition, the fitted model specification replicated the structure of the population model. For all the equivalence tests, this is a power condition. In other words, the population fit value (RMSEA or CFI) falls within the pre-determined EB (e.g., .05 for the unmodified RMSEA). This is also a power condition for both ICTs, since, as described above, the population fit value falls within the EB (although recall that the ICTs are not formal inferential tests, and therefore do not have a formal definition for power or Type I error). The χ^2 goodness of fit test should retain its null hypothesis of perfect fit; therefore, any rejection of the null is a Type I error.

All equivalence tests in the negligible misspecification condition are power conditions because the population CFI is greater than .95 and the population RMSEA value is less than .05 (population values are available under Figures 1, 2, and 3). This is also a power condition for both ICTs for the same reason. In contrast, the χ^2 goodness of fit test should reject the null hypothesis H_0: $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$, because the null indicates that the model's structure perfectly represents the structure of the data (which it does not, because a negligible misspecification is still a misspecification).

In the non-negligible misspecification condition, where the fit value is at or beyond the equivalence interval, the χ^2 goodness of fit test should once again reject the null hypothesis, reflecting a power condition. For the equivalence testing procedures, none of the equivalence tests are expected to reject their corresponding null hypotheses for both modified and unmodified EBs because the population CFI is less than or equal to .95 and the population RMSEA value is

greater than or equal to .05; rejecting these null hypotheses is a Type I error. The same is true for both ICTs.

In order to select the best procedures, tests had to have satisfactory Type I error control at the EB. Further, the best-performing tests also had to have good relative power and Type I error control beyond the EB.

The nominal Type I error rate (α) was set to .05 for all conditions. We conducted simulations until there were 1000 converged model solutions for each condition (non-converged solutions were discarded). Having 1000 replications corresponds to a standard error in reported proportions of approximately .007 (when the true proportion is equal to .05, i.e., the fit index is set at the EB). The number of bootstrap samples for the YHY procedures was set at 500. For the number of bootstrap samples, 500 resamples corresponds to a standard error in means of approximately .0008 for both CFI and RMSEA.

Results

Tables 2 to 5 summarize the results of the simulation study. The results of the non-negligible misspecification at the EB for all six scenarios defined above (CFI and RMSEA for each of the three models) are in Table 2. The CFI ICT is presented for the CFI scenario and the RMSEA ICT is presented for the RMSEA scenario. Tables 3 to 5 each pertain to one of the three models (Table 3 = Model 1, Table 4 = Model 2, Table 5 = Model 3). The remaining misspecification conditions (non-negligible beyond the equivalence interval, negligible misspecification, and perfect fit) are displayed in each of the three tables as different rows for every unique sample size.

For each of the three models, the number of non-converged solutions was higher with small sample sizes. This pattern of non-convergence was also mimicked for the bootstrap

samples. Specifically, across the models and misspecification types displayed in Tables 3 to 5, we obtained the following ranges of incidences of non-convergence: $N = 50$: 7-231, $N = 75$: 1-117, $N = 100$: 0-91, $N = 200$: 0-21, $N = 400$: 0-2; $N \geq 800$: 0.

In the sections below, we summarize the results of each test for each condition, beginning with misspecifications where the relevant bound of CI was aligned with the associated EB.

Non-Negligible Misspecification at the Equivalence Bound

In each of these scenarios, well-performing equivalence tests should have Type I error rates near .05. To decide whether Type I error rates for a given test are accurate within a specific scenario, we use Bradley's (1978) liberal criterion ($\alpha \pm .5\alpha$; in this case .025 to .075).

Table 2 depicts the Type I error control of the 10 equivalence tests in conditions where the underlying model fit is meant to be as close as possible to .95 for CFI and .05 for RMSEA. Once again, the corresponding CFI ICT results are reported for the CFI scenario, whereas the RMSEA ICT results are reported for the RMSEA scenario (recall that there is no formal hypothesis tested with the ICTs).

In the CFI scenarios, across all three models, the $EBF\alpha_{CFI}$ and the EBF_{CFI} displayed inconsistent Type I error control. With the largest sample size of 5000, the $EBF\alpha_{CFI}$ Type I error rates ranged from .003 (for Model 1) to .041 (for Model 3). The EBF_{CFI} Type I error control was also variable, ranging between .024 for Models 1 and 2 to .079 for Model 3 (for $N = 5000$). In contrast, the $EBFB_{CFI}$ test demonstrated the best Type I error control: it was the only CFI test whose performance at $N = 5000$ fell within Bradley's (1978) liberal criterion for all three models, falling between .043 and .045.

CFI tests with modified EBs had poor Type I error control. Given a sample size of 5000, Type I errors rates ranged from .070 to .244 for Model 1, .138 to .333 for Model 2, and .211 to

.284 for Model 3. This finding is expected, given that the underlying EB has been modified for these tests (and is lower than .95 in most cases).

Unlike the CFI tests, both RMSEA tests with unmodified EBs had similar Type I error performance. Both the $EBFB_{RMSEA}$ and the EBF_{RMSEA} demonstrated good Type I error control, falling between .041 to .067, within Bradley's (1978) criterion. RMSEA tests with modified EBs had poorer Type I error control than those with unmodified EBs. Specifically, given a sample size of 5000, across the three models, Type I error rates ranged from .180 (Model 1; $EBF_{RMSEA\text{-}A}$) to .468 (Model 2; $EBF_{RMSEA\text{-}A}$). Once again, this finding is expected given that the underlying EB for these tests has been modified.

The ICTs are not formal inferential tests; therefore, we would not hypothesize any specific Type I error control (which is another drawback to using them as pseudo-hypothesis tests). Instead, the ICTs in these scenarios correspond to the proportion of instances where the RMSEA estimate is less than .05 or the CFI estimate is greater than .95. Note that given larger sample sizes, for tests with good Type I error control ($EBFB_{CFI}$, EBF_{RMSEA}, and $EBFB_{RMSEA}$), the ICTs have values close to .50.

Next, we present results for the remaining three misspecification conditions, starting with the non-negligible condition beyond the EB.

Non-Negligible Misspecification Beyond the Equivalence Bound

The misspecification in this set of conditions is beyond the bound of the equivalence interval for all equivalence tests, so any rejection of the null hypothesis (i.e., H_0: $RMSEA_{pop} \geq EB_{upper}$ or H_0: $CFI_{pop} \leq EB_{lower}$) is a Type I error. For the unmodified EBs, the CFI equivalence tests had Type I error rates as high as .047 (i.e., the $EBFB_{CFI}$ test for $N = 100$ in Model 1). The EBF_{CFI} test and $EBF\alpha_{CFI}$ test had comparable error rates at all sample sizes (the largest

discrepancy occurred in Model 3 with a sample size of 200 where the error rates of the EBF_{CFI} and the $EBF\alpha_{CFI}$ were .032 and .014, respectively). Error rates for all unmodified EB CFI tests rise slightly and then fall as the sample size increases, approaching 0 at the largest sample size of 5000. This finding also occurred for the $EBFB_{CFI\text{-}A}$, the $EBF_{CFI\text{-}A}$, and $EBF\alpha_{CFI}$ tests in all models. The error rates for the modified EB tests were much higher than their unmodified counterparts, going as high as .865 for the $EBFB_{CFI\text{-}A}$ test in the $N = 75$ condition in Model 1. Generally, the error rates for the modified EBs varied widely but remained high for some conditions (e.g., Model 1, $N = 1000$, error rate = .584 for the $EBFB_{CFI\text{-}A}$ test) until $N = 5000$, where they all approached 0.

For unmodified EBs, there was almost no difference between the results of the $EBFB_{RMSEA}$ and the EBF_{RMSEA} tests, as the Type I error rates were both at or near zero for all sample sizes and all models. In contrast, for modified EBs, the $EBF_{RMSEA\text{-}A}$ test had slightly higher error rates than the $EBFB_{RMSEA\text{-}A}$ test at smaller sample sizes (e.g., in Model 1, for $N = 75$, the error rate for the $EBF_{RMSEA\text{-}A}$ test was .095 and was .061 for the $EBFB_{RMSEA\text{-}A}$ bootstrap test). Error rates for RMSEA equivalence tests sometimes displayed a similar pattern to the CFI equivalence tests, where they would rise and then fall slightly with increasing sample size.

This pattern of error rates rising and then falling with increasing sample sizes results from widely variable CFI/RMSEA values at smaller sample sizes. The rising rates within smaller sample sizes occurs because there is a greater probability of rejecting the null hypothesis with greater N (e.g., $N = 100$ vs. $N = 50$). As sample sizes become larger (e.g., $N > 200$), there is less variability in CFI/RMSEA values and error rates decrease in turn.

The Type I error rates for the ICTs were considerably higher than the error rates for equivalence tests with unmodified EBs. This result is expected because the CI used in each given

equivalence test is structured around the fit index itself, making the magnitude of the fit index higher (for CFI) or lower (for RMSEA) than the relevant bound of the CI used for comparison. Error rates were near-zero at a sample size of 5000 for Models 2 and 3. Error rates reached 0 for the RMSEA ICT at $N = 400$ for Model 1. The CFI ICT did not reach 0 for Model 1 even at a sample size of 5000 (the CFI ICT error rate for this condition was .044). Thus, the ICTs suggest evidence for good fitting models more frequently than the equivalence tests, especially in smaller samples (however, recall that the ICTs are not formal inferential tests and do not exhibit any formal Type I error control, as demonstrated in the previous section).

Comparing the ICTs to equivalence tests with modified EBs, the RMSEA ICT results were relatively similar to the $EBFB_{RMSEA\text{-}A}$ results with $N \geq 100$ and near-identical to the $EBF_{RMSEA\text{-}A}$ across models and sample sizes. In contrast, the CFI equivalence tests with the modified EBs tended to have larger error rates than the CFI ICTs.

Because this condition imposes model misspecification, the χ^2 goodness of fit test should reject the null hypothesis. The goodness of fit test in the Monte Carlo study rejected the null hypothesis in all instances with $N \geq 400$ for all models but failed to reject the null hypothesis of perfect fit in almost 50% of replications for $N = 50$ for Model 1 (and 23.2% and 36.0% of replications in Models 2 and 3, respectively).

In comparing the results of the equivalence tests of unmodified and modified bounds (as well as the ICTs), we caution that further comparisons between the procedures, namely power comparisons, should consider that many of the tests had poor Type I control (specifically, all ICTs, all tests with modified EBs, as well as the EBF_{CFI} and $EBF\alpha_{CFI}$).

Negligible Misspecification

Negligible misspecification is a power condition for all equivalence tests. Generally, as

expected, both ICTs and equivalence tests with modified EBs had higher power than equivalence tests with unmodified EBs. However, note that both ICTs and equivalence tests with modified EBs also performed poorly in the previously described non-negligible conditions.

At lower sample sizes, the EBF_{CFI} test had greater power than the $EBF\alpha_{CFI}$ test. For example, in Model 3 in the $N = 100$ condition, the EBF_{CFI} test had a power rate of .540 compared to .422 for the $EBF\alpha_{CFI}$ test. The power of the $EBFB_{CFI}$ equivalence test was lower at small sample sizes than both the $EBF\alpha_{CFI}$ test and the EBF_{CFI} tests, but became similar to or greater than power of the $EBF\alpha_{CFI}$ test and the EBF_{CFI} tests at $N = 200$ for all models (note, however, that in the non-negligible misspecification condition at the EB, the $EBFB_{CFI}$ test demonstrated superior Type I error control relative to both the $EBF\alpha_{CFI}$ and the EBF_{CFI}). All CFI-based equivalence tests for unmodified EBs approached power = 1.00 by a sample size of 400. CFI equivalence tests using modified EBs had, as expected, higher power than their unmodified counterparts and approached power = 1.00 by $N = 200$.

Given smaller sample sizes, RMSEA-based equivalence tests with an analytic CI computation outperformed the bootstrap tests. Once again, for unmodified EBs, the EBF_{RMSEA} test had more power than the $EBFB_{RMSEA}$ test. Given modified EBs, the $EBF_{RMSEA\text{-}A}$ test had more power than the $EBFB_{RMSEA\text{-}A}$ test. Power became largely similar for both tests at $N = 400$ for both modified and unmodified EBs. The RMSEA equivalence tests for unmodified EBs required large sample sizes to reach the power = 1.00 (i.e., power was above .900 by $N = 1000$ for Models 2 and 3 and $N = 5000$ for Model 1). The power of the RMSEA equivalence tests using modified EBs was considerably better, reaching high power at sample sizes of 200 (e.g., Model 3, $N = 200$, $EBFB_{RMSEA\text{-}A}$ and $EBF_{RMSEA\text{-}A}$ power rates were .888 and .911, respectively).

The ICTs had higher power than the equivalence tests at lower sample sizes. The RMSEA ICT became comparable with RMSEA equivalence tests using unmodified EBs at N = 800 for Models 2 and 3 and at N = 5000 for Model 1. The CFI ICT became comparable with CFI equivalence tests using unmodified EBs at N = 400.

It is noteworthy that the results for the modified EB tests and the ICT were very similar. Namely, the RMSEA ICT was nearly identical to the $EBF_{RMSEA\text{-}A}$ test for all sample sizes. By N = 400, it was also comparable to the $EBFB_{RMSEA\text{-}A}$. The CFI ICT was comparable to all equivalence tests with modified EBs by N = 200.

Once again, as in the non-negligible misspecification condition outside of the EB, the χ^2 goodness of fit test should reject the null hypothesis because the covariance matrix implied by the (misspecified) hypothesized theoretical model is not equal to the data-generating (population) covariance matrix. However, there were many occurrences at small to moderate sample sizes where this test failed to reject the null hypothesis of perfect fit (e.g., for Model 1 at N = 200, the χ^2 goodness of fit test had a Type II error rate of 71.9%). As expected, power of this test increase with increasing sample sizes, where at the highest sample size of N = 5000, all null hypotheses are rejected. It is noteworthy that even though this misspecification is within the equivalence interval and therefore deemed negligible in equivalence testing, because of the hypothesis structure of the goodness of fit test even small misspecifications should be rejected.

Perfect Fit

Perfect fit is a power condition for all of the equivalence tests. Once again, as in the negligible misspecification condition, ICTs and equivalence tests with modified EBs were more powerful than equivalence tests with unmodified EBs. These results should not be interpreted

without also considering that ICTs and equivalence tests with modified EBs exhibited poorer Type I error control than equivalence tests with unmodified EBs.

For the CFI-based equivalence tests (both with unmodified and modified EBs), as expected, tests with a 100(1 – 2α)% CI had higher power to detect improved fit than those with 100(1 – α)% CI, especially at lower sample sizes. For instance, in Model 1 with $N = 200$, the EBF_{CFI} test had a power = .972 whereas the $EBF\alpha_{CFI}$ power was .924. Power rates became comparable between the EBF_{CFI} test and the $EBF\alpha_{CFI}$ test for all models at $N = 400$, where they reached a power = 1.00. For modified EBs, the $EBF_{CFI\text{-}A}$ and $EBF\alpha_{CFI\text{-}A}$ were comparable at $N = 100$. The $EBFB_{CFI}$ had power near 0 at the smallest sample size ($N = 50$) but power increased to at least .305 with $N = 100$ for all models. The $EBFB_{CFI\text{-}A}$ had higher power, even with samples as low as 75, for all models. For modified EBs, power for all three CFI equivalence tests was consistently high for all three models, where even at a sample size of 75, the lowest power rate for all three CFI tests was .905 (this rate belonged to the $EBFB_{CFI\text{-}A}$ in Model 2, power rates for the rest of the CFI tests ranged from .925 to .999).

The RMSEA-based equivalence tests with an analytic computation of the CI had higher power than the bootstrap CI tests at lower sample sizes for all models (namely, for unmodified EBs, the EBF_{RMSEA} test had higher power than the $EBFB_{RMSEA}$ test; for modified EBs, the $EBF_{RMSEA\text{-}A}$ test had higher power than the $EBFB_{RMSEA\text{-}A}$ test). For example, in Model 3, at $N = 100$, power for the EBF_{RMSEA} and $EBFB_{RMSEA}$ was .186 and .000, respectively. In the same condition for modified EB tests, power was .792 for $EBF_{RMSEA\text{-}A}$ and .631 for $EBFB_{RMSEA\text{-}A}$. Power became similar for these tests at $N \geq 400$ for unmodified EBs and $N \geq 200$ for modified EBs. Power approached 1 for RMSEA equivalence tests with unmodified EBs with $N \geq 800$ for

all three models. RMSEA equivalence tests with modified EBs, as expected, approached power = 1 with smaller sample sizes than their unmodified counterparts (at $N = 400$).

As expected, power was considerably higher for the ICTs than for the equivalence tests with unmodified EBs at lower sample sizes. For example, in Model 2, given a sample size of 75, power for the CFI ICT was .912, while power for the CFI equivalence tests with unmodified EBs ranged from .126 ($EBFB_{CFI}$) to .222 (EBF_{CFI}). In the same condition, the RMSEA ICT had a power rate of .669, while RMSEA equivalence tests with unmodified EBs had power ranging from 0 ($EBFB_{RMSEA}$) to .099 (EBF_{RMSEA}). Power rates became relatively comparable between the ICTs and all CFI equivalence tests at $N \geq 200$. Specifically, at a sample size of 200, power reached 1 for all CFI ICTs while the power ranges for all CFI equivalence tests were .924 - .991, .967 – 1.00, .991 - .997, for Models 1, 2, and 3, respectively. Power rates became comparable between the ICTs and the RMSEA equivalence tests with $N \geq 400$ (except the EBF_{RMSEA} in Model 1 having power = .772 compared to the ICTs for RMSEA power = .999).

Like in the negligible misspecification condition, the ICTs and modified EB tests performed similarly. The RMSEA ICT had nearly identical power to the $EBF_{RMSEA\text{-}A}$ at all sample sizes across all models. The RMSEA ICT results were also comparable to the bootstrap test results for all models with $N \geq 400$. The CFI ICT results were comparable to CFI equivalence tests with modified EBs with $N \geq 100$ for all models.

Recall that the χ^2 goodness of fit test should fail to reject its null hypothesis in the perfect fit condition. The smaller the sample size, the more frequently the χ^2 goodness of fit test rejected the perfect fit hypothesis for all three models. At $N = 50$, rejections of the null hypothesis were as high as .313 for Model 2. As the sample size increased, Type I error rates were near 5%, as expected, because the nominal Type I error probability was set to $\alpha = .05$.

Illustrative Example

To illustrate how equivalence-based fit tests may be used in conjunction with traditional fit measures for SEM, we present an example of a confirmatory factor analysis (CFA) using real data. The data for the CFA was from an interactive online version of the Humor Styles Questionnaire, which contained 32 questions adapted from Martin et al. (2003), each with a response scale ranging from 1 (*never or very rarely true*) to 5 (*very often or always true*). Of interest was how well the proposed four-factor structure (representing constructs named *affiliative*, *self-enhancing*, *aggressive*, and *self-defeating*) fit the sample data. The dataset consisted of $N = 993$ participants (https://openpsychometrics.org/_rawdata/). See Figure 5 for the proposed CFA model structure. We assume a nominal Type I error rate (α) of .05.

First, we used the `lavaan` package (Rosseel, 2012) to fit the model and create a model object (see https://osf.io/uwk6m for full example code). Next, we obtained descriptive fit indices using the `fitmeasures()` function in `lavaan`. We used the `neg.cfi()` and `neg.rmsea()` functions to conduct equivalence tests for the CFI and RMSEA indexes; we also used these functions to supplement the reporting of the bootstrap equivalence tests with the average of the bootstrap samples. Both of these functions can be accessed in the `negligible` package (Cribbie et al., 2023).

The RMSEA, CFI, SRMR for the hypothesized four factor model were .060, .842 and .067, respectively. The null hypothesis of the chi-square goodness of fit test was rejected, $\chi^2(458) = 2088.81$, $p < .001$.

Of the six CFI equivalence tests that are compared in this manuscript, three use modified EBs ($EBF\alpha_{CFI\text{-}A}$, $EBF_{CFI\text{-}A}$, and $EBFB_{CFI\text{-}A}$) and three do not ($EBF\alpha_{CFI}$, EBF_{CFI}, and $EBFB_{CFI}$). The three variations of CI_{lower} are compared to an EB (modified or unmodified). The CI types are

the Yuan et al. (2016) α/2-tail lower bound 95% CI, our proposed modified α-tail lower bound 90% CI, and the α-tail lower bound 90% YHY bootstrap CI (with 1000 bootstrap samples). In this example, the three CI_{lower} values are 0.818 for the Yuan lower bound, 0.822 for the modified lower bound, and 0.819 for the YHY bootstrap lower bound.

First, we discuss the findings from the CFI tests with unmodified EBs. Comparing the lower bounds of the CIs to the common equivalence interval bound for CFI (0.95), for all three equivalence tests the lower bound of the CI_{lower} for CFI is less than the EB (i.e., EBF_{CFI}: $0.822 \leq 0.95$; $EBF\alpha_{CFI}$: $0.818 \leq 0.95$; $EBFB_{CFI}$: $0.819 \leq 0.95$). Therefore, for all three tests (EBF_{CFI}, $EBF\alpha_{CFI}$, and $EBFB_{CFI}$), we fail to reject the hypothesis H_0: $CFI_{pop} \leq EB_{lower}$, indicating that these methods would lead to a conclusion that the hypothesized CFA model for the Humor Styles Questionnaire may have a non-negligible misspecification (i.e., not enough evidence to reject the null hypothesis).

For Yuan et al.'s (2016) modified EBs, the CFI = .95 boundary is adapted to a CFI = .940 (see OSF example at https://osf.io/uwk6m for how to obtain the modified bound). Comparing the CI_{lower} to the modified equivalence interval bound, we obtain the same conclusions (i.e., $EBF_{CFI\text{-}A}$: $0.822 \leq 0.940$; $EBF\alpha_{CFI\text{-}A}$: $0.818 \leq 0.940$; $EBFB_{CFI\text{-}A}$: $0.819 \leq 0.940$). Once again, we fail to reject the null hypothesis (H_0: $CFI_{pop} \leq EB_{lower}$) that the population model fit falls at or beyond the EB ($EBF_{CFI\text{-}A}$, $EBF\alpha_{CFI\text{-}A}$, and $EBFB_{CFI\text{-}A}$).

Of the four equivalence tests for RMSEA, two use modified EBs ($EBF_{RMSEA\text{-}A}$ and $EBFB_{RMSEA\text{-}A}$) and two do not (EBF_{RMSEA} and $EBFB_{RMSEA}$). Two variations of CI_{upper} can be obtained for equivalence tests for RMSEA. They are the MacCallum et al. (1996) α upper bound 90% CI and the YHY bootstrap α upper bound 90% CI (with 1000 bootstrap samples). The two upper bounds of each CI are .062 and .065, respectively.

Again, we first demonstrate the conclusions of the equivalence tests using unmodified EBs. Comparing the CI_{upper} to the unmodified equivalence interval bound for RMSEA (0.05), the upper bound of the CI for RMSEA is greater than the EB in both scenarios (namely, EBF_{RMSEA}: .062 ≥ .05; $EBFB_{RMSEA}$: .065 ≥ .05). Therefore, for both tests (EBF_{RMSEA} and $EBFB_{RMSEA}$), we fail to reject the null hypothesis H_0: $RMSEA_{pop} \geq EB_{upper}$, again indicating that the hypothesized four-factor model may have a non-negligible misspecification.

For the two tests using modified EBs ($EBF_{RMSEA\text{-}A}$ and $EBFB_{RMSEA\text{-}A}$), the RMSEA = .05 boundary is adapted to an RMSEA of .056 (details regarding the computation of the adapted RMSEA can be found in the example on OSF). Comparing the upper bound of the CI to the modified EB, we again fail to reject the null hypothesis for both procedures (i.e., $EBF_{RMSEA\text{-}A}$: .062 ≥ .056; $EBFB_{RMSEA\text{-}A}$: .065 ≥ .056), also indicating the model may have a non-negligible model misspecification.

The model fit results for the four-factor CFA may be written in a manuscript as follows (we use the $EBFB_{CFI}$ and $EBFB_{RMSEA}$ for the sample write-up since they were the best performing in the simulation study; we are assuming that the author had defined these acronyms within their paper):

> The estimated four-factor CFA model for humour styles did not fit the data well. The RMSEA, CFI, TLI, and SRMR for the model are .060, .842, .829, and .067. The $EBFB_{CFI}$ test (average CFI_B = .838, 90% CI_{lower} = 0.819) failed to reject H_0: $CFI_{pop} \leq EB_{lower}$. The $EBFB_{RMSEA}$ test (average $RMSEA_B$ = .061, 90% CI_{upper} = .065) also failed to reject H_0: $RMSEA_{pop} \geq EB_{upper}$, indicating that the degree of model misfit may be non-negligible.

We can see that the equivalence-based fit tests provide hypothesis testing information to supplement the values of the traditional fit measures. In this example, the model fit is generally

poor so the fit measures, along with the inferential tests, all provide some evidence towards a similar conclusion. The χ^2 goodness of fit tests rejects the hypothesis of perfect fit and the equivalence tests fail to find evidence of negligible misspecification. However, as demonstrated in the simulation study, there are situations where the information drawn from inferential tests and descriptive measures is not as consistent. Regardless of the specific scenario, the use of equivalence tests contributes important inferential information regarding model fit.

Discussion

Before researchers interpret the parameter estimates of an SEM model, they should first assess the fit of the model to the data. This assessment is often completed using a variety of model fit indices. Often, the only inferential test accompanying the model fit indices is the χ^2 goodness of fit test, which has an unrealistic null hypothesis of perfect fit. The fit indices themselves usually become utilized as ICTs, with values being compared to cut-offs, instead of being used on a more descriptive and stand-alone basis (as was originally intended).

The introduction of equivalence testing into the SEM literature by MacCallum et al. (1996) and Yuan et al. (2016) has allowed for inclusion of inferential tests with the process of evaluating and reporting model fit beyond the χ^2 goodness of fit test, facilitating conclusions about whether the degree of model misfit is negligible. We hope that the incorporation of equivalence tests into model fit reporting leads to a decrease in reliance on using fit indices as ICTs.

This study proposed a modification to Yuan et al.'s (2016) original equivalence test for the CFI, termed the EBF_{CFI}. We also proposed the $EBFB_{RMSEA}$ and $EBFB_{CFI}$, which rely upon the YHY bootstrap procedure. We compared the performance of these tests, using both modified and unmodified EBs, to ICTs and the traditional χ^2 goodness of fit test. We found that our proposed

equivalence tests ($EBFB_{RMSEA}$ and $EBFB_{CFI}$) along with the EBF_{RMSEA} generally performed well, having high power at medium to large sample sizes, as well as adequate Type I error control. Finally, we demonstrated how the results from the equivalence tests of model fit may be interpreted (in tandem with other fit measures) using an illustrative example.

Selecting an appropriate cut-off value for an EB is difficult. In our Monte Carlo simulation study, we used EBs of .95 for CFI and .05 for RMSEA because these comparison values are extremely popular among SEM researchers; however, it is important for researchers to consider what the minimum meaningful amount of misfit should be given the research context when setting the EBs. Yuan et al. (2016) proposed modified EBs for use in equivalence testing for evaluating model fit. In our Monte Carlo study, we found that these modified EBs had higher power than their original counterparts, but also higher Type I error rates. Counsell et al. (2020) came to a similar conclusion regarding the use of adjusted EBs for RMSEA. Recent research has introduced a simulation-based method for "dynamic fit index" cut-offs that can be customized to a specific CFA model (McNeish & Wolf, 2021). These dynamic cut-off values might be applied as EBs for a specific model in question, but more research is needed to investigate how they may be used and how these modified equivalence tests perform.

It is important to note that ICTs will always have greater (pseudo-) power and (pseudo-) Type I error rates than equivalence tests because the parameter (e.g., CFI, RMSEA) estimate, rather than the bound of the CI, is being compared to the EB. Of note, because they are not formal inferential tests, ICTs have no interpretational advantage relative to equivalence testing procedures. All that is gained from performing ICTs is a quick comparison of a fit index to a reference point. Fit indices can offer more to model fit evaluation by being interpreted as effect sizes (e.g., extent of model fit improvement relative to a null model) instead of as ICTs. We hope

that the incorporation of equivalence testing in SEM will dissuade researchers from using fit indices as ICTs.

Out of the three CFI tests, we recommend the $EBFB_{CFI}$. In our Monte Carlo study, we found that the $EBFB_{CFI}$ was similar to the other two tests in terms of power and possessed the best Type I error control when the underlying model fit was at the equivalence interval. An additional advantage of the $EBFB_{CFI}$ is that it uses the traditional 100(1 - 2α)% CI for comparison with the EB, unlike the original $EBF\alpha_{CFI}$ test which has a 100(1 - α)% CI for comparison. With respect to the RMSEA, both the EBF_{RMSEA} and the $EBFB_{RMSEA}$ tests perform similarly well across model conditions. We recommend the use of either for interpreting or reporting model fit. All of these tests can be performed using the `neg.cfi()` and `neg.rmsea()` functions within the `negligible` package (Cribbie et al., 2023).

The χ^2 goodness of fit test performed poorly with any kind of model misspecification (negligible or non-negligible) in our Monte Carlo study. Given the hypotheses of this test, it is logical that any time the model does not perfectly replicate the features of the data, the hypothesis of perfect fit will be rejected. It is unsurprising, therefore, that the use of the χ^2 goodness of fit test on its own is not recommended (see Browne & Cudeck, 1993; Tanaka, 1987). However, without equivalence tests, the test is often the sole inferential piece of evidence provided for assessing model fit. In accordance with past research, we do not recommend the use of this test, and suggest that greater emphasis be placed on equivalence testing for model fit evaluation instead.

There are a couple of limitations in this study to highlight. First, as with all Monte Carlo studies, the results are only directly relevant to the models and conditions investigated in this paper. For example, we used one of many methods for simulating model misspecification

(consistent with Chen et al., 2008), but other methods of simulating misspecification are possible. However, we expect that the general patterns and differences between the various equivalence tests, ICTs, and the χ^2 goodness of fit test will generalize to other kinds of models, forms of model misspecification, and sample sizes. Second, we only used data that was simulated from a multivariate normal distribution; therefore, we did not investigate how the equivalence tests perform when this assumption is violated. Notably, Lai (2019) has proposed an analytic confidence interval for CFI when normality is violated, but it has yet to be incorporated into an equivalence testing context.

Equivalence tests may be useful for other fit indices beyond RMSEA and CFI, since there are numerous fit indices available for evaluating SEM models. Accordingly, future directions may explore equivalence tests for other fit indices such as the standardized root mean residual (SRMR) or Tucker-Lewis Index (TLI).

This study aimed to contribute to the SEM and equivalence testing literature by providing additional information for investigating the fit of SEMs. Equivalence-based inferential tests of model fit can provide valuable evidence to supplement descriptive fit indices.

CHAPTER THREE

EQUIVALENCE TESTING BASED FIT INDEX: STANDARDIZED ROOT MEAN SQUARED RESIDUAL (STUDY 2)

Structural equation modeling (SEM) can be useful for assessing how multiple variables (both observed and latent) are related to one another. Hypothesized associations between variables can be specified and estimated in an SEM model. However, before these associations are interpreted, it is necessary to ascertain how well the researcher-hypothesized SEM model fits the data. Model fit can be assessed in numerous ways.

One of the most common ways to investigate model fit relies upon goodness of fit indices. There are various indices that evaluate model fit descriptively according to different criteria: the root mean squared error of approximation (RMSEA: Steiger & Lind, 1980) evaluates the amount of discrepancy error per the specified model's degree of freedom; the comparative fit index (CFI: Bentler, 1990) compares the fit of the observed model to a null model with no associations among the variables; and the standardized root mean squared residual (SRMR: Bentler, 1995) measures the mean value of residual correlations among observed variables after the model has been fit to the data.

When first introduced into SEM, descriptive fit indices were to be interpreted as effect sizes, allowing the researcher to evaluate the fit of their model along a continuous spectrum (McNeish & Wolf, 2021). In practice, fit indices have come to be used as quasi-significance tests or informal check tests (ICTs), where their value is compared to a pre-determined cut-off to reach a binary decision about whether the model fits the data well or not. This practice is problematic for numerous reasons. First, ICTs exhibit no formal Type I error control because

they are not inferential tests. Further, they cloud the use of fit indices by making them solely pieces of information for ICTs, instead of effect sizes.

To disentangle the descriptive function of fit indices from ICTs, it may be useful to introduce an inferential test into the evaluation of model fit, which can be accomplished with the use of equivalence testing. Equivalence testing, or negligible effect testing, allows researchers to evaluate a lack of association among variables. The equivalence testing approach has already been introduced into the framework of SEM through equivalence tests for the RMSEA and CFI (MacCallum et al., 1996; Yuan et al., 2016; Study 1). However, there has yet to be an equivalence test for the SRMR, despite it being one of the three most-reported fit indices (with the RMSEA and CFI being the other two; Jackson et al., 2009).

Although inferential tests are secondary in their importance to interpreting the degree of misfit from a fit index itself, they can still add useful insights to model fit evaluation. For this reason, the present study adds to the literature in three parts. First, we propose different variations of equivalence tests based on the SRMR. Next, we use a Monte Carlo simulation study to compare the performance of these tests to one another, to ICTs, and to the χ^2 goodness of fit test, which is the standard inferential test reported in SEM. Finally, we use an illustrative example with real data to demonstrate how equivalence tests for SRMR may be incorporated into model fit evaluation and reporting.

The Standardized Root Mean Squared Residual and its Confidence Interval

When a model is specified in an SEM framework, it implies a covariance structure for the observed variables. This covariance structure, called the model-implied covariance matrix, is compared, based on some criterion, to a covariance matrix that results from the data. Specifically, the estimation of model parameters involves the minimization of a discrepancy

function with respect to the model parameters $\boldsymbol{\theta}$. Under the assumption of multivariate normality, the following discrepancy function is minimized for maximum likelihood (ML) estimation:

$$F_{\text{ML}} = \ln\left|\widehat{\boldsymbol{\Sigma}}\right| - \ln|\mathbf{S}| + \text{tr}\left(\mathbf{S}\,\widehat{\boldsymbol{\Sigma}}^{-1}\right) - p\ ,$$

where $\widehat{\boldsymbol{\Sigma}}$ is the model-implied covariance matrix, $\mathbf{S}$ is the covariance matrix for the sample data, p is the number of manifest (observed) variables, | | is the determinant of a matrix, and tr() is the trace function. Alternatively, for different estimators such as unweighted least squares (ULS) or the asymptotically distribution free (ADF) weighted least squares, the following discrepancy function is minimized:

$$F = (\mathbf{s} - \boldsymbol{\sigma}_0)'\,\hat{\mathbf{W}}\,(\mathbf{s} - \boldsymbol{\sigma}_0)\ ,$$

where $\boldsymbol{\sigma}_0 = \boldsymbol{\sigma}(\boldsymbol{\theta}_0)$ is the vectorized model-implied covariance structure, $\boldsymbol{\sigma}$ is a $t = p(p + 1)/2$ vector of population covariances of p observed variables, and the elements in $\boldsymbol{\theta}_0$ are the parameters that must be estimated from the sample data of length $q \leq t$ (Maydeu-Olivares, 2017). The term $\mathbf{s}$ refers to the vector of sample covariances of length t, while $\hat{\mathbf{W}}$ is a weight matrix with $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$; with an increasing number of observed variables, the estimate of the weight matrix converges towards a population weight matrix**.** The choice of different matrices for $\hat{\mathbf{W}}$ leads to different estimators (ULS, ADF, and so on).

SRMR

The SRMR fit index is based on the residual covariance matrix, $\mathbf{S} - \widehat{\boldsymbol{\Sigma}}$. Once the residual covariance matrix is standardized, the average of the residual elements is calculated. When $\mathbf{S}$ and $\widehat{\boldsymbol{\Sigma}}$ are identical, F is equal to 0. Similarly, if the specified model fits the data well, the residual covariance vectors, $\mathbf{s} - \boldsymbol{\sigma}_0$, will be small and SRMR will be near zero. The formula for the population SRMR is

$$SRMR_{pop} = \sqrt{\frac{(\boldsymbol{\sigma} - \boldsymbol{\sigma_0})' \, \mathbf{G^{-1}} (\boldsymbol{\sigma} - \boldsymbol{\sigma_0})}{t}} = \sqrt{\frac{1}{t} \sum_{i \leq j} \left(\frac{\sigma_{ij} - \sigma_{ij}^0}{\sqrt{\sigma_{ii}\sigma_{jj}}} \right)^2}$$

where **G** is a diagonal matrix with elements $\sigma_{ii}\sigma_{jj}$.

As noted by Maydeu-Olivares (2017), the formula for the original sample SRMR (SRMR: Bentler, 1995) is,

$$\widehat{SRMR}_{\mathbf{B}} = \sqrt{\frac{(\mathbf{s} - \hat{\boldsymbol{\sigma}})' \, \hat{\mathbf{G}}^{-1} \, (\mathbf{s} - \hat{\boldsymbol{\sigma}})}{t}} = \sqrt{\frac{1}{t} \sum_{i \leq j} \left(\frac{s_{ij} - \hat{\sigma}_{ij}}{\sqrt{s_{ii}s_{jj}}} \right)^2}$$

where $\hat{\mathbf{G}}$ can be defined as a diagonal matrix containing elements $s_{ii}s_{jj}$. Maydeu-Olivares (2017) demonstrated that this sample SRMR overestimates the population SRMR. This bias is also exacerbated at low sample sizes and low factor loadings (Shi et al., 2018).

Maydeu-Olivares (2017) derived a formula for an asymptotically unbiased estimate of SRMR (irrespective of any discrepancy function and assumptions regarding the distribution) as

$$\widehat{SRMR}_U = \hat{k}_s^{-1} \sqrt{\frac{\max(\mathbf{e_s}'\mathbf{e_s} - \mathrm{tr}(\hat{\boldsymbol{\Xi}}_s), 0)}{t}} \text{ , with}$$

$$\hat{k}_s = 1 - \frac{\mathrm{tr}\left(\hat{\boldsymbol{\Xi}}_s^{\,2}\right) + 2\mathbf{e_s}'\hat{\boldsymbol{\Xi}}_s\mathbf{e_s}}{4(\mathbf{e_s}'\mathbf{e_s})^2},$$

where $\mathbf{e_s}$ is a vector of length t which contains all of the standardized residual covariances from matrix $\hat{\boldsymbol{\Xi}}_s$. Matrix $\hat{\boldsymbol{\Xi}}_s$ is the estimate of the asymptotic covariance matrix of $\mathbf{e}_s$.

A confidence interval (CI) for the sample $SRMR_U$ can be estimated using a reference normal distribution and asymptotic standard errors (Maydeu-Olivares et al., 2017):

$$CI_{lower} = \widehat{SRMR}_U - z_{\alpha/2} SE(\widehat{SRMR}_U)$$

and

$$CI_{upper} = \widehat{SRMR}_U + z_{\alpha/2} SE(\widehat{SRMR}_U),$$

with

$$SE(\widehat{SRMR}_U) = \sqrt{k_s^{-2} \frac{\mathrm{tr}\left(\widehat{\boldsymbol{\Xi}}_s^{\,2}\right) + 2\mathbf{e}_s' \widehat{\boldsymbol{\Xi}}_s \mathbf{e}_s}{2t\, \mathbf{e}_s' \mathbf{e}_s}} .$$

Equivalence Testing

Originally applied to biopharmaceutical disciplines, equivalence testing was introduced to social science research by articles such as Tryon (2001) and Rogers et al. (1993). Equivalence testing is appropriate when a researcher would like to infer a lack of association. For instance, rather than aiming to confirm that the means of two groups are different, a researcher may instead hypothesize that the difference between the two means is so small that it may be considered negligible or unimportant. This hypothesis can never be directly tested with traditional null hypothesis significance testing (NHST). Specifically, traditional NHST may either provide evidence to infer a difference between the group means or fail to do so. In other words, traditional NHST cannot serve as evidence that two populations have means that are equivalent to one another (see Rogers et al., 1993).

This limitation directly applies to the case of inference for model fit in SEM. In SEM, the most common inferential test reported is the χ^2 goodness of fit test with null hypothesis

$$H_0: \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}},$$

where $\boldsymbol{\Sigma}$ represents the population covariance matrix of the variables and $\hat{\boldsymbol{\Sigma}}$ is the model-implied population covariance (Bollen, 1989).

Like the example of group means presented above, a researcher can never directly infer that the two population covariance matrices are equal to one another using the χ^2 goodness of fit test. It is therefore no surprise that this test has been previously criticized. Researchers have noted that this null hypothesis is unrealistic because any statistical model is necessarily an over-simplification of reality. If the null hypothesis is false to any extent (i.e., the model fit is not perfect), a large sample size will lead to an increased likelihood of its rejection (Bentler & Bonett, 1980; Browne & Cudeck, 1992; Steiger, 2007; Hooper et al., 2008).

Even though there are notable limitations to the χ^2 goodness of fit test, it is often the sole piece of statistical inference reported for evaluating model fit in SEM. However, just as equivalence tests have been proposed for comparing means (Rogers et al., 1993; Wellek, 2010), so too have there been equivalence tests proposed for model fit in SEM (i.e., McCallum et al., 1996; Yuan et al., 2016; Study 1). These additions can serve to supplement descriptive fit indices like SRMR and add inferential information beyond that of the χ^2 goodness of fit test.

Equivalence Tests for SRMR

In the SEM context, equivalence tests compare the fit of a model to a cutoff that represents the best fit value that is considered intolerable. This cutoff is synonymous with the minimally meaningful effect size estimate (or MMES) in equivalence testing research (see, for example Yuan et al., 2016). In equivalence testing research, the MMES is used to set values for an equivalence interval. Similarly, in a fit index for SEM setting, the MMES is approximated by a single value, which is one bound of an equivalence interval (we refer to this value as an equivalence bound, EB).

To create an equivalence test for SRMR, we use the logic that has been previously developed in creating equivalence tests for RMSEA and CFI (see MacCallum et al., 1996; Yuan et al., 2016; Study 1). Recall that higher values of SRMR are indicative of poorer fit (corresponding to the average value of the elements in the residual matrix, $\mathbf{S} - \hat{\boldsymbol{\Sigma}}$, being large). Accordingly, any SRMR estimate larger than (or equal to) the value of the upper EB (EB_{upper}) would be too large to be tolerable. Because RMSEA is also a lower-is-better fit index, equivalence tests for SRMR share a common structure with equivalence tests already developed for RMSEA. Thus, the null hypothesis of an equivalence test for SRMR can be expressed as

$$H_0: SRMR_{pop} \geq EB_{upper} ,$$

and may be successfully rejected when the level of a model's misspecification is tolerable. The rejection of the null hypothesis provides support for the alternate hypothesis,

$$H_1: SRMR_{pop} < EB_{upper} .$$

To test the null hypothesis, equivalence tests compare the bounds of CIs to EBs. Because equivalence tests for fit indices are one sided, it is only necessary to assess one bound of the CI and compare it to the corresponding EB (e.g., CI_{upper} is compared to EB_{upper} for SRMR). Structuring the test in this way allows for a nominal Type I error rate of α. Note that this structure is still consistent with traditional equivalence testing, where a symmetric $100(1 - 2\alpha)\%$ CI is compared against the lower and upper bounds of the equivalence interval. In this setting, the probability of a Type I error rate in a given tail is α.

Proposed Modifications to Equivalence Tests for SRMR

The section above presents a general framework for an equivalence test for SRMR. However, modifications to these equivalence testing procedures are proposed. These

modifications are applicable to the computation of the CI and the choice of the minimally tolerable size of misspecification (the value of EB_{upper}).

Bootstrapped Confidence Intervals. Although Maydeu-Olivares et al. (2017) derived a method for obtaining CIs for $SRMR_U$ using a reference normal distribution and asymptotic standard errors, another version of an equivalence test for SRMR may use a CI that is obtained via bootstrapping. This procedure would be possible when either $SRMR_U$ or $SRMR_B$ is the fit index of interest. A bootstrap approach uses repeated sampling from a given dataset with replacement. The estimate of interest (e.g., an SRMR value) is computed at every sampling iteration. Finally, the distribution of estimates is used to obtain a CI by selecting the appropriate quantiles (e.g., 95^{th} percentile).

The current study also proposes an equivalence testing procedure for SRMR that employs the Yuan et al. (2007) bootstrap, which is subsequently referred to as the YHY bootstrap. To ensure that the noncentrality parameters (which capture the level of model misfit) between the bootstrap data and the full sample data match as closely as possible, the YHY bootstrap applies a data transformation before beginning the bootstrap process (Zhang & Savalei, 2016). The transformation ensures that the noncentrality parameters capturing the level of model misfit between the bootstrap data and the full sample data match one another as closely as possible. The YHY bootstrap was found to perform better than the naïve bootstrap for SRMR in prior research; however, both methods only produced good coverage rates in certain scenarios (see Zhang & Savalei, 2016 for more information). Zhang and Savalei (2016) note that the existence of bias in the estimate of SRMR may contribute to the poorer performance of YHY bootstrap (relative to the good coverage rates of RMSEA and CFI in the same study). Accordingly, it may be of interest to combine the YHY procedure using the $SRMR_U$ estimate into a bootstrap-based

equivalence test. If bootstrap tests for fit indices are used, estimates of the fit index itself may be presented as the average value of the bootstrap estimates (i.e., average $SRMR_U$ or average $SRMR_B$), rather than calculated using the sample SRMR formula presented above.

Modified and Unmodified Equivalence Bounds. Traditional cut-off values that have been applied to SRMR as ICTs are a potential option for identifying an EB value for an SRMR equivalence test. Indeed, these have been the values used for EBs in equivalence tests for RMSEA and CFI (see MacCallum et al., 1996 or Yuan et al., 2016). In an often-cited simulation study exploring how well different cut-off values are able to discriminate between correctly specified and incorrectly specified models (in terms of ICTs), Hu and Bentler (1999) found that a value of .08 resulted in a reasonable ratio of Type I error rates to power for SRMR. A cut-off value of .05 was also explored by Hu and Bentler and was found to be more conservative, rejecting a higher proportion of true-population models. Both values (or other similar values) could be used as options for EB_{upper}.

However, using traditional cut-off values for EB_{upper} may have limitations. Commonly used cut-offs for ICTs do not generalize to different kinds of models (see Marsh et al., 2004; McNeish & Hancock. 2018). The same may be true when the cut-offs are used as EBs. Further, as noted by Shi et al. (2018), the Hu and Bentler (1999) study utilized the original biased sample estimate derived for SRMR ($SRMR_B$) rather than $SRMR_U$, the unbiased sample estimate later derived by Maydeu-Olivares (2017). Rather than using an ICT value of .08, Shi et al. (2018) identified closely fitting models by comparing $SRMR_U$ at the population level to $.05 \times \overline{R}^2$, where $\overline{R}^2$ is average communality of the observed variables (this comparison was one component of a two-index strategy, with the second component involving the magnitude of the largest absolute value in a matrix of standardized residual covariances). The modified cut-off value was able to

detect model misspecification when an SEM model's factor structure was misspecified and when an SEM model's cross-loadings were omitted. However, the modified cut-off did not perform well when there were omitted covariances between residuals (Shi et al., 2018). These modified cut-offs were recommended over an ICT value of .08. The proposed cut-offs for good and acceptable fit were SRMR $\leq .05 \times \overline{R}^2$ and SRMR $\leq .10 \times \overline{R}^2$, respectively (Shi et al., 2022).

An equivalence test using a modified EB has the same hypothesis structure as an equivalence test with an unmodified EB. Namely,

$$H_0: \text{SRMR}_{\text{pop}} \geq \text{modified EB}_{\text{upper}}$$

and

$$H_1: \text{SRMR}_{\text{pop}} < \text{modified EB}_{\text{upper}}$$

would represent the null and alternate hypotheses for equivalence tests with modified EBs, respectively.

It is unclear how either cut-off, modified or unmodified, would perform as an EB for equivalence testing. For the remainder of the manuscript, we refer to an SRMR cut-off of .05 or .08 as an unmodified EB and refer to a cut-off of $.05 \times \overline{R}^2$ or $.10 \times \overline{R}^2$ as a modified EB.

The Present Study

Together with RMSEA and CFI, SRMR is one of the three most common fit indices reported (Jackson et al., 2009). However, there has yet to be an equivalence test pertaining to SRMR proposed, even though equivalence tests have been developed for both RMSEA and CFI (MacCallum et al., 1996; Yuan et al., 2016). The present study defines and compares a set of equivalence tests for SRMR. Accordingly, we introduce tests structured around both SRMR_{B} and SRMR_{U}, modified and unmodified EBs, along with CIs that are computed analytically or via

bootstrap. Further, we compare the performance of these novel tests to traditional methods of evaluating model fit in SEM such as ICTs and the χ^2 goodness of fit test.

Monte Carlo Simulation Study

To compare the proposed SRMR equivalence tests with ICTs and the χ^2 goodness of fit test, we conducted a Monte Carlo simulation study using R (R Core Team, 2021). There were several equivalence test variations: Some used modified EBs and some did not, some were based on $SRMR_B$ while others were based on $SRMR_U$, and some used a bootstrapped CI while others used an analytically derived CI. Across all of these combinations, there were twelve total equivalence tests proposed and investigated. Table 6 presents their names and definitions.

The equivalence tests, ICTs, and the χ^2 goodness of fit test all assess different hypotheses, whereas all SRMR equivalence tests share the same hypotheses. The SRMR tests evaluate the research (alternate) hypothesis that the population SRMR value falls below the EB_{upper} (e.g., .05 and .08 for unmodified EBs or $.05 \times \overline{R}^2$ and $.10 \times \overline{R}$ the modified EBs). The ICTs test whether an SRMR value (either $SRMR_U$ or $SRMR_B$) is smaller than a given cut-off value (one of .05, .08, $.05 \times \overline{R}^2$, or $.10 \times \overline{R}$). Despite the ICTs not being a formal inferential testing method, we incorporate this approach into our simulation study because of its extremely common use. Finally, the χ^2 goodness of fit test evaluates the research (alternate) hypothesis of whether there are any deviations between the population covariance matrix and the model-implied covariance matrix.

To run the simulation, we used both the `simsem` (Pornprasertmanit et al., 2020) and `SimDesign` (Chalmers & Adkins, 2020) packages. Using `SimDesign` allowed for the establishment of a simulation structure for the project, while `simsem` allowed for enhanced data analysis capabilities from SEM Monte Carlo studies. The start values for model estimation were

the defaults of the `lavaan` package (Rosseel, 2012). Specifically, latent variances started at 0.05, observed and residual variances started at half of the corresponding observed sample variance, regression coefficient start values were computed using ordinary least squares, covariances started at 0, and factor loading start values used the FABIN3 (instrumental variables) method.

Population Generating Models

The path diagrams for the models, including misspecifications that will be outlined in greater detail below, are in Figures 6 to 11. The population generating models are displayed in Figures 6, 8, and 10, while the misspecified (beyond the EB) versions of these models are specified in Figures 7, 9, and 11. These models were adapted from Chen et al. (2008), who originally generated the models by using a combination of their own expertise and experience with common SEM models found in the social science literature. To create the four different misspecification types we utilized this study -- namely (1) perfect fit, (2) negligible misspecification, (3) non-negligible misspecification beyond the EB, and (4) non-negligible misspecification at the EB -- the models' factor loadings were changed slightly from those used by Chen et al. Models 1, 2, and 3 all had three latent variables. The models had a varying number of indicators, some of which cross-loaded onto multiple latent variables. Model 1 had nine indicator variables with each latent variable measured by four indicators. Model 2 had fifteen indicators with each latent variable measured by six indicators. Model 3 had nine indicators with each latent variable measured by four indicators. Additionally, Model 3 had four correlated exogenous variables as regressors for the latent variables.

We defined factor loading parameters to match our criteria for creating misspecification. Once misspecified models were fit to samples drawn from the population generating models, they satisfied criteria for both the negligible and non-negligible conditions (i.e., both $SRMR_B$ and

$SRMR_U < .05 \times \overline{R}^2$ for the negligible condition and both $SRMR_B$ and $SRMR_U > .08$ for the non-negligible condition outside of the EB). Models 1, 2, and 3 had primary standardized factor loadings of 0.70. For Model 1, indicators that had a non-zero association with more than one latent variable had a cross-loading value of 0.32. For Model 2, cross-loading values were 0.25. For Model 3, cross-loading values were 0.21. Figures 6 to 11 also include the population values of the fit indices for the misspecified models.

Test Specifications

The unmodified EBs were set at either .05 or .08. We also explored how well the cut-offs proposed by Shi et al. (2018) would function as modified EBs. Accordingly, we also computed two adapted EBs using $.05 \times \overline{R}^2$ and $.10 \times \overline{R}^2$ as EBs. By fitting the models to corresponding population covariance matrices, we were able to create both power and Type I error conditions. To create power conditions, we confirmed that population values of SRMR were below $.05 \times \overline{R}^2$. Similarly, to create the non-negligible misspecification conditions outside of the EB to investigate Type I error rates (more detail on all misspecification conditions is provided below), we confirmed that population values of SRMR were above .08.

Manipulated Factors for the Monte Carlo Study

The two primary factors manipulated in our Monte Carlo study were sample size and degree of model misspecification. Following Chen et al. (2008), we included $N = 50$, 75, 100, 200, 400, 800, 1000, and 5000. These sample sizes represent those commonly found in social and behavioural science research.

Four types of misspecified models were fit to samples from each of the three population generating models. These misspecifications include: (1) perfect fit, (2) negligible misspecification, (3) non-negligible misspecification beyond the EB, and (4) non-negligible

misspecification at the EB. For the perfect fit condition, the correctly specified model was fit to samples drawn from each of the three population generating models. For the negligible misspecification condition, each of the models had the following omissions: Model 1 omitted a cross-loading from latent variable 2 to indicator 7; Model 2 omitted a cross-loading from indicator 11 to latent variable 2; and Model 3 omitted two cross-loadings (one from indicator 7 to latent variable 2 and another from indicator 6 to latent variable 3). For the non-negligible misspecification beyond the EB, the models were misspecified as follows: Model 1 instead fit a one-factor model over all of the indicator variables (rather than a three-factor model); Model 2 instead fit a two-factor model over the indicator variables (rather than a three-factor model) such that indicators 1 to 5 and 13 mapped onto one latent variable and indicators 6 to 12, 14, and 15 mapped onto another latent variable; and Model 3 instead fit a two-factor model over the indicator variables (rather than a three-factor model) such that indicators 1 to 5 had free loadings on latent variable 1 and indicators 6 to 9 had free loadings on latent variable 2. For the non-negligible misspecification beyond the EB, Model 3 also omitted a regression coefficient (between exogenous variable 3 and factor 2); further, because factor 3 no longer existed, the previously existing paths between two of the exogenous variables and this factor were also eliminated.

In the non-negligible misspecification condition at the EB, we created two scenarios to investigate the equivalence tests' Type I error performance. We simulated one scenario using an unmodified EB of .05, and another scenario using a modified EB of $.10 \times \overline{R}^2$. In each Type I error scenario, we generated conditions so that the underlying model fit was as close as possible to the EB. Accordingly, for the first scenario, both SRMR indices were as close as possible to .05. For the second scenario, both SRMR indices were as close as possible to $.10 \times \overline{R}^2$ (the

specific value differed for each of the three models because the $\overline{R}$ differed across the models). To create these scenarios, we modified the fitted models used in the negligible misspecification condition. The modification involved fixing one of the coefficient values in the fitted model to a different magnitude than the population model instead of allowing it to be estimated freely. For each scenario, the models were fit to the corresponding population covariance matrices to confirm that the SRMR indices were exactly .05 or .10 × $\overline{R}^2$ to three decimal places.

Power and Error Conditions

Figure 12 illustrates the locations of the population values of the fit indices relative to the EBs across the misspecification conditions. In the perfect fit condition, the structure of the fitted model was the same as the structure of the population model. This is a power condition for all equivalence tests because the population SRMR value of 0 is below the EB. Because the population SRMR is below the EB, this condition also functions as a power condition for ICTs, despite ICTs not having formal definitions of both power and Type I error rate. Conversely, the χ^2 goodness of fit test should retain the null hypothesis that the population covariance matrix is identical to the model-implied covariance matrix. Accordingly, a rejection of the null hypothesis for the χ^2 goodness of fit test is a Type I error in the perfect fit condition.

In the negligible misspecification condition, the population SRMR is less than .05 × $\overline{R}^2$, making it a power condition for all equivalence tests. This feature also makes this condition function as a power condition for all ICTs. In contrast, the null hypothesis of the χ^2 goodness of fit test should be rejected since the population covariance matrix and the model-implied covariance matrix are different.

For the non-negligible misspecification conditions, where the SRMR value is either at the EB or beyond the EB, all equivalence tests should fail to reject their null hypotheses. This

expectation is because the population SRMR is either at or larger than the EB; any rejection of the null hypothesis is a Type I error. This situation also applies to the ICTs. The χ^2 goodness of fit test should also fail to reject the null hypothesis because, like in the negligible misspecification condition, the population covariance matrix and the model-implied covariance matrix are not the same.

In order to select the best equivalence tests, the procedures had to demonstrate satisfactory Type I error control at the EB. The best-performing procedures also had to have good relative power and Type I error rates beyond the EB.

The Monte Carlo simulation continued replications until there were 1000 converged model solutions for every condition (non-converged solutions were discarded). There were 500 bootstraps for equivalence tests employing the YHY procedure for computing a CI. The nominal Type I error rate (α) was set to .05 for all conditions. With 1000 replications, the associated standard error in reported proportions is approximately 0.007 (when the true proportion is equal to .05, i.e., the fit index is set at the EB). For the number of bootstrap samples, 500 resamples corresponds to a standard error in the means of approximately 0.0005 and 0.0004 for $SRMR_U$ and $SRMR_B$, respectively.

Results

The results of the Monte Carlo study are in Tables 7 to 14. Tables 7 and 8 present results for the two scenarios in which the degree of misspecification was such that the population SRMR was equal to the EB (Table 7 contains results for the models when the EB was set to .05; Table 8 contains results for the models when the EB was set to .10 × $\overline{R}^2$). Tables 9 to 14 present results for the other misspecification conditions for all three models (each model is summarized

in two tables: Tables 9 and 10 for Model 1; Tables 11 and 12 for Model 2; Tables 13 and 14 for Model 3).

Across all three models, non-convergence rates decreased as sample size increased. For models and misspecification presented in Tables 9 to 14, the following ranges of incidences of non-convergence occurred: $N = 50$: 2-235; $N = 75$: 0-81; $N = 100$: 0-57; $N = 200$: 0-4; $N \geq 400$: 0. Bootstrap samples also had the same pattern of non-convergence, where the highest incidence of non-convergence occurred for the smallest sample sizes.

Non-Negligible Misspecification at the Equivalence Bound

When testing non-negligible misspecification at the EB, equivalence tests that perform well should have Type I error rates that are close to .05. Bradley's (1978) liberal criterion was used to determine whether Type I error rates were acceptable within a given scenario (the acceptable range is $\alpha \pm .5\alpha$; or .025 to .075 for the present study).

Results for the population SRMR = .05 scenario are in Table 7. For completeness, we present results for all SRMR equivalence tests as well as the two ICTs that compare the observed SRMR (either biased or unbiased) directly to .05. Results for the population SRMR = $.10 \times \overline{R}^2$ scenario are in Table 8. In this scenario, the magnitude of the modified EB value changed depending on the model but ranged from .052 (Model 2) to .057 (Model 1).

Scenario 1: Equivalence Bound of .05

With population SRMR = .05, at sample sizes ≥ 400 the $ESRMR_{U05}$ was the only test that had acceptable Type I error control out of the three equivalence tests that had an EB of .05 ($ESRMR_{U05}$, $ESRMRY_{U05}$, and $ESRMRY_{05}$). At $N = 5000$, error rates for this test ranged from .038 (Model 2) to .048 (Model 1). Tests with a bootstrap CI ($ESRMRY_{U05}$ and $ESRMRY_{05}$) tended to have Type I error rates outside of the upper bound of the Bradley criterion.

Given that the other nine equivalence tests have different underlying EBs, we would not expect their Type I error rates to be near .05. Accordingly, all tests with an EB of .08 or EB of .10 × $\overline{R}^2$ had higher Type I error rates that were outside of the Bradley criterion. As expected, tests with an EB of .05 × $\overline{R}^2$ had lower error rates than .05 that were also outside of the range of the liberal criterion.

Scenario 2: Equivalence Bound of .10 Multiplied by Average of Communalities

With population SRMR = .10 × $\overline{R}^2$, at sample sizes ≥ 400, only the $ESRMR_{U10A}$ had an acceptable Type I error rate out of the three equivalence tests that had an EB of .10 × $\overline{R}^2$ ($ESRMR_{U10A}$, $ESRMRY_{U10A}$, and $ESRMRY_{10A}$). At N = 5000, error rates for the $ESRMR_{U10A}$ ranged from .032 (Model 1) to .039 (Model 2). Like in the first scenario, tests with a bootstrap CI (this time $ESRMRY_{U05A}$ and $ESRMRY_{05A}$) tended to have rates beyond the upper bound of Bradley's criterion.

Similar to the first scenario, the nine other equivalence tests have EBs that are different from .10 × $\overline{R}^2$; therefore, their Type I error rates are not expected to be close to .05. Thus, as anticipated, all tests with an EB of .08 had higher Type I error rates, outside of the upper bound of the Bradley criterion. Tests with an EB of .05 or .05 × $\overline{R}^2$ had error rates that were lower than the lower bound of the Bradley liberal criterion.

Informal Check Tests

An additional consequence of using ICTs as hypothesis tests is that they are not anticipated to exhibit any specific Type I error control, because they are not formal inferential tests. Instead, they illustrate the proportion of instances where the SRMR estimate (biased or unbiased, depending on the ICT in question) is less than .05 (Scenario 1) or less than .10 × $\overline{R}^2$ (Scenario 2). In both scenarios, the ICT based on the unbiased estimate of the SRMR was much

closer to a proportion of .5 than the ICT based on the biased estimate of SRMR. This pattern was especially prevalent at smaller sample sizes.

Non-Negligible Misspecification Beyond the Equivalence Bound

For non-negligible misspecifications beyond the EB, a rejection of the null hypothesis (H_0: $SRMR_{pop} \geq EB_{upper}$) is a Type I error because the population value of SRMR is larger than EB_{upper}**.** Error rates were higher with smaller sample sizes; however, all equivalence tests had low Type I error rates overall with rates under 5% across all tests and models. With $N \geq 400$, all equivalence tests had rejection rates of zero, with most tests' rejection rates being at or near-zero with sample sizes as low as 50. The largest Type I error rate observed was .033 for Models 1 and 2 and .031 for Model 3 (for all models, these rates pertained to the $ESRMRY_{U08}$ test).

At $N = 50$ and $N = 75$, equivalence tests with modified EBs (both $.05 \times \overline{R}^2$ and $.10 \times \overline{R}^2$) tended to have lower error rates than equivalence tests with unmodified EBs of .08. At these sample sizes, error rates of modified EB tests were similar to unmodified EB tests of .05. This result is logical in that multiplying either .05 or .10 by the average communality estimate shrinks the resulting EB for the modified EB tests, leading to a lower likelihood of rejecting the null hypothesis.

Tests based on $SRMR_U$ tended to have somewhat larger error rates than tests based on $SRMR_B$. This effect was more apparent for unmodified EBs and dampened when tests with modified EBs were used. For example, in Model 2, with $N = 75$, the error rates for the $ESRMR_{U08}$ and the $ESRMRY_{U08}$ were .013 and .033, respectively, but were 0 for $ESRMRY_{08}$. For modified EB tests in the same scenario (Model 2 and $N = 75$), all rates were 0.

Although error rates were low across all tests and models, the highest error rate tended to belong to the $ESRMRY_{U08}$ test. This result is logical because this test contains the largest of the

EBs compared in the simulation study (EB = .08), which allows for more models to be classified as well-fitting. In the $ESRMRY_{U08}$ test, this EB was used in tandem with a bootstrap CI around $SRMR_U$, leading to the SRMR equivalence test with the highest error rate.

As expected, ICTs tended to have error rates that were similar or higher than error rates for the equivalence tests. For example, in Model 3, given a sample size of 50, the range of error rates for all equivalence tests was between .000 and .031, whereas the range for ICTs in the same condition was between .000 and .297. All ICTs reached error rates of near zero (.001 or less) at a sample size of 800. ICTs using .08 as a comparison value, as expected, tended to have higher error rates than ICTs using other comparison values. ICTs using the $SRMR_U$ tended to have higher Type I error rates than ICTs using $SRMR_B$. For this reason, the $SRMR_U$ ICT with .08 as a comparison had the highest error rate across all models.

Recall that the null hypothesis for the χ^2 goodness of fit test states that the model-implied covariance matrix equals the population covariance matrix. Accordingly, this test should reject the null hypothesis in this condition. Indeed, the χ^2 goodness of fit test rejected all null hypotheses with $N \geq 100$. However, there were non-zero Type II error rates at lower sample sizes (e.g., retentions went as high as 2.5% in in Model 1).

The next two sections of the results outline power rates for all equivalence tests, pseudo-power rates for the ICTs, and power rates and Type I error rates for negligible and perfect misspecifications, respectively, for the χ^2 goodness of fit test. When making comparing these procedures, keep in mind that the ICTs had no specific Type I error control in the non-negligible misspecification at the EB condition. Equivalence tests with a bootstrap CI also tended to have poorer Type I error control than tests with an analytic computation of the CI.

Negligible Misspecification

The negligible misspecification condition is a power condition for all equivalence tests in this study. All equivalence tests reached a power of 1.00 or near 1.00 by $N = 5000$. For both kinds of EBs (modified and unmodified), power varied depending on the specific type of EB selected (i.e., either .05 × $\overline{R}^2$ or .10 × $\overline{R}^2$ for modified EB tests or .05 or .08 for unmodified EB tests). Even though we describe these differences below, the description is not meant to serve as a direct comparison between the tests because it is logical that a test with a lower EB value would be associated with higher power. Equivalence tests with modified EBs of .10 × $\overline{R}^2$ had rejection rates of 1.00 or near 1.00 (i.e., > .995) by $N = 400$. In contrast, equivalence tests with modified EBs of .05 × $\overline{R}^2$ performed much worse, only reaching 1.00 or near 1.00 (> .996) at $N = 5000$. As expected, equivalence tests with unmodified EBs = .08 performed best, reaching a power of 1.00 or near 1.00 by $N = 200$ (where all tests were at 1.00 except $ESRMR_{U08}$ in Model 1, which had a power = .998). Tests with unmodified EBs of .05 reached a power of 1.00 at $N = 800$, having inferior power to equivalence tests with .08 and .10 × $\overline{R}^2$ EBs but performing better than equivalence tests with modified EBs = .05 × $\overline{R}^2$.

Across all models and at all sample sizes before power rates reached a ceiling, tests based on $SRMR_U$ tended to have higher power than tests based on $SRMR_B$. For example, in Model 3 with $N = 800$, $ESRMRY_{U05A}$ and $ESRMR_{U05A}$ had power rates of .926 and .756, respectively, while $ESRMRY_{05A}$ had a considerably lower power rate of .395. This pattern was evident in equivalence tests employing both modified and unmodified EBs.

In many instances across the Monte Carlo study, there was low power at smaller sample sizes for tests with CIs constructed using the YHY bootstrap. This result was especially noticeable when bootstrap CIs were constructed around the $SRMR_B$ index. For example, in Model 2, power rates only became greater than zero for the $ESRMRY_{05}$ test with $N \geq 200$; for

the same model, power rates only became greater than zero for the $ESRMRY_{05A}$ test with $N \geq$ 800). Tests employing the YHY bootstrap procedure for the CI became relatively comparable to tests using an analytic computation for $N = 400$ for unmodified EBs. For modified EBs, tests with bootstrap and analytic CIs were only comparable at $N = 5000$.

Across all models, the $ESRMRY_{U08}$ test reached power = 1.00 the quickest of all equivalence tests. However, the $ESRMR_{U08}$ test had higher power than the $ESRMRY_{U08}$ test with $N = 50$ and, when sample size increased (i.e., $N = 200$ or greater), was comparable to the $ESRMRY_{U08}$ test, but with slightly lower power. With $N \geq 400$, the $ESRMRY_{U10A}$ and the $ESRMR_{U10A}$ tests also had excellent power, comparable to both the $ESRMR_{U08}$ and $ESRMRY_{U08}$ tests.

Power for ICTs was quite variable in the negligible condition. There was a pronounced difference between ICTs using $SRMR_U$ and ICTs using $SRMR_B$, where ICTs based on $SRMR_U$ had much higher power than their $SRMR_B$ counterparts. This pattern was quite evident at small sample sizes (e.g., in Model 2, given a sample size of 50, power rates for $SRMR_U$ and $SRMR_B$ ICTs using a comparison value of .05 × $\overline{R}^2$ were .611 and .000, respectively). The $SRMR_U$ ICT with a .08 comparison had the highest power rate of all ICTs. Most ICTs approached power = 1.00 by $N = 800$, except the .05 × $\overline{R}^2$ ICTs (both for $SRMR_U$ and $SRMR_B$), which did not reach power = 1.00 until $N = 5000$. Generally, the highest power ICTs tended to have higher power than equivalence tests at smaller sample sizes. However, across all models, both the $ESRMR_{U08}$ and $ESRMRY_{U08}$ tests had comparable power to the highest power ICTs with a sample size as low as 100.

In the negligible misspecification condition, even though this misspecification is negligible for all equivalence tests and within the EB, the null hypothesis for the χ^2 goodness of

fit is false in this condition. However, the test often resulted in Type II errors, even at moderate sample sizes (e.g., in Model 1 with $N = 200$, this test was nonsignificant in 73.0% of replications). Rejections were more common as sample size increased and every null hypothesis was successfully rejected with $N = 5000$.

Perfect Fit

Perfect fit is a power condition for all equivalence tests because the population SRMR is smaller than EB_{upper}. For all three models, most equivalence tests reached power = 1.00 at $N = 1000$ (having power > .993).

Equivalence tests with unmodified EBs tended to reach power = 1.00 at the same rate or quicker than equivalence tests with modified EBs. Like in the negligible condition, the specific type of EB chosen for modified ($.10 \times \overline{R}^2$ and $.05 \times \overline{R}^2$) and unmodified (.08 and .05) tests had notable effects on power. Like in the negligible condition, this finding is not meant to be a direct comparison between the tests themselves but can help illustrate the expected effects different EBs have on power rates. For all models, equivalence tests with unmodified EBs of .08 and .05 reached power = 1.00 at $N \geq 200$ and $N \geq 400$, respectively. For modified EB tests, tests using $.10 \times \overline{R}^2$ and $.05 \times \overline{R}^2$ as their EBs reached power near 1.00 at $N \geq 400$ and $N \geq 1000$ (with all tests having power > .993), respectively.

Like in the negligible misspecification condition, tests built around the $SRMR_U$ index tended to have higher power than tests based upon $SRMR_B$. Similar to the negligible condition, this pattern was evident for equivalence tests using both modified and unmodified EBs. For example, in Model 2, given $N = 200$, the $ESRMR_{U05}$ and the $ESRMRY_{U05}$ tests had power = .987 and 1.00, respectively, compared to the $ESRMRY_{05}$, which had power = .345 in the same condition. Given the same model and sample size presented, the $ESRMR_{U10A}$ test and the

$ESRMRY_{U10A}$ test had power = .994 and 1.00, respectively, while the $ESRMRY_{10A}$ test had power = .726.

Similar to the negligible condition, tests using a YHY bootstrap CI sometimes had power near 0 at the lowest sample sizes, especially when CIs were constructed around the $SRMR_B$. Within a given EB (i.e., .05, .08, .10 × $\overline{R}^2$, and .05 × $\overline{R}^2$), tests using a bootstrap method for calculating a CI often only became comparable to a test using an analytic computation of a CI when all three tests (i.e., $SRMR_U$ test with bootstrapped CI, $SRMR_B$ test with bootstrapped CI, and $SRMR_U$ test with analytic CI) had a sufficient sample size for all of them to attain power = 1.00.

Like in the negligible misspecification condition, the $ESRMRY_{U08}$ test tended to reach power = 1.00 the quickest of all equivalence tests being compared. The $ESRMR_{U08}$ test consistently had the highest power rates with $N = 50$. The $ESRMR_{U08}$ test also had relatively comparable power to the $ESRMRY_{U08}$ test across all samples (the largest disparity occurred at $N = 75$ in Model 1, where the $ESRMRY_{U08}$ test and the $ESRMR_{U08}$ test had power rates of .949 and .897, respectively). With $N \geq 200$, the $ESRMRY_{U10A}$ and $ESRMR_{U10A}$ tests also displayed comparable power to both the $ESRMRY_{U08}$ test and the $ESRMR_{U08}$ test.

Like in the negligible misspecification condition, power was quite variable across ICTs in the perfect fit condition. Once again, there was a pronounced difference between ICTs that were based on $SRMR_U$ compared to ICTs based on the $SRMR_B$ such that $SRMR_U$ ICTS had higher power than $SRMR_B$ counterparts. For instance, in Model 3 at $N = 400$, given a comparison value of .05 × $\overline{R}^2$, $SRMR_U$ ICT had power = .999, while the $SRMR_B$ ICT had power = .800. Most ICTs approached power = 1.00 with $N \geq 400$, except for .05 × $\overline{R}^2$ ICTS for both $SRMR_U$ and $SRMR_B$, which reached power = 1.00 with $N \geq 800$ and $N \geq 1000$, respectively. Although equivalence

tests tended to have lower power than the best performing ICTs at small sample sizes, once again, the $ESRMR_{U08}$ and $ESRMRY_{U08}$ tests had comparable power rates to the best performing ICTs, starting from sample sizes as low as 75 (for Models 2 and 3) and 100 (Model 1).

In the perfect fit condition, the χ^2 goodness of fit should retain the null hypothesis. Across all models, the test rejected the null hypothesis (i.e., Type I error) most often at smaller sample sizes. With increasing sample size, the Type I error rate decreased, but did not reach zero. Because α was set at .05, the χ^2 goodness of fit maintained incorrect rejections around 5% for the highest sample sizes.

Illustrative Example

The following example demonstrates how an SEM model may be evaluated using equivalence-based fit tests for SRMR. The example is based on a confirmatory factor analysis (CFA) model estimated by Reifler et al. (2011) evaluating British Foreign Policy opinions using national surveys collected between May and September of 2008. The authors found foreign policy beliefs in contemporary Britain were governed by two factors which were labeled British Militarism and Liberal Internationalism (see Reifler et al., 2011, for more information). The CFA model proposed by the authors is illustrated in Figure 13.

The example described here is a reanalysis of openly accessible data (see Reifler et al., 2018) used in the Reifler et al. (2011) article. The sample size was $N = 6{,}155$, with 4,871 participants remaining after removing cases with missingness. Full annotated code for the example is openly available on OSF (see https://osf.io/dq8sa/?view_only=df38766c227742d5b60e5d79b099df82). This example is meant to exhibit how the proposed equivalence tests may be used in practice; thus, there are some small differences between the analyses presented here and the analysis in the original article (e.g., we

use ML estimation whereas the original authors used Weighted Least Squares Mean and Variance (WLSMV) estimation, we did not impute missing data, and so on). We used a nominal Type I error rate (α) of .05.

To estimate the model, we used the `lavaan` package (Rosseel, 2012). The RMSEA, CFI, and $SRMR_U$ for the model were .038, .991, and .014, respectively. Equivalence tests for SRMR were performed using the `neg.srmr()` function, which we created and is available on OSF (https://osf.io/dq8sa/?view_only=df38766c227742d5b60e5d79b099df82). For completeness, we present results for all 12 SRMR equivalence tests proposed and compared in this manuscript. In practice, usually only one SRMR equivalence test would be conducted and reported; however, for example, some researchers may be interested in using the SRMR with different EB values.

Six equivalence tests use modified EBs and six use original EBs. The modified EBs are $.05 \times \overline{R}^2$ and $.10 \times \overline{R}^2$. The unmodified EBs are .05 and .08. The upper bound of the CI used for equivalence testing can be structured around either $SRMR_U$ or $SRMR_B$. A YHY bootstrapped CI is possible for both fit indices (here, the number of bootstrap samples was set to 1000), while an analytic computation of a CI is possible for $SRMR_U$ only. The upper bounds of the $SRMR_U$ analytic CI, the $SRMR_U$ YHY bootstrap CI, and the $SRMR_B$ YHY bootstrap CI were .017, .018, and .019, respectively.

Beginning with an unmodified EB of .05, we can compare the upper bounds of the CIs to the EB for all three equivalence tests ($ESRMR_{U05}$, $ESRMRY_{U05}$, and $ESRMRY_{05}$). For all three equivalence tests, the upper bound of the CI for SRMR is smaller than the EB (namely, $ESRMR_{U05}$: .017 < .05; $ESRMRY_{U05}$: .018 < .05; $ESRMRY_{05}$: .019 < .05). Therefore, for all three tests, we can reject the null hypothesis that H_0: $SRMR_{pop} \geq EB_{upper}$ and conclude that the model fits the data (i.e., the degree of misfit is negligible).

These conclusions do not change with an unmodified EB of .08. Specifically, the three equivalence tests (this time, $ESRMR_{U08}$, $ESRMRY_{U08}$, and $ESRMRY_{08}$) all have an upper bound of a CI that is smaller than the EB ($ESRMR_{U08}$: .017 < .08; $ESRMRY_{U08}$: .018 < .08; $ESRMRY_{08}$: .019 < .08). Once again, the null hypothesis (H_0: $SRMR_{pop} \geq EB_{upper}$) can be rejected for all three tests.

A modified EB of .05 × $\overline{R}^2$ corresponds to .0181 (details of this computation are in the OSF file). In this scenario, two of the three equivalence tests ($ESRMR_{U05A}$ and $ESRMRY_{U05A}$) reject the null hypothesis (H_0: $SRMR_{pop} \geq EB_{upper}$). For $ESRMR_{U05A}$: .017 < .0181 and for $ESRMRY_{05A}$: .018 < .0181. The $ESRMRY_{05A}$ fails to reject its null hypothesis because .019 ≥ .0181.

Finally, a modified EB of .10 × $\overline{R}^2$ corresponds to .0361. All three corresponding equivalence tests ($ESRMR_{U10A}$, $ESRMRY_{U10A}$, and $ESRMRY_{10A}$) reject the null hypothesis H_0: $SRMR_{pop} \geq EB_{upper}$ ($ESRMR_{U10A}$: .017 < .0361; $ESRMRY_{U10A}$: .018 < .0361; $ESRMRY_{10A}$: .019 < .0361).

As the $ESRMR_{U08}$ was the best performing equivalence test in the study, we provide these results in the write up of model fit for the CFA below (we have assumed that the authors would have defined the $ESRMR_{U08}$ acronym earlier in their manuscript):

> The estimated two-factor CFA model evaluating British Foreign Policy opinions had excellent fit to the data. The RMSEA, CFI, $SRMR_U$, and TLI for the model were .038, .991, .014, and .986. The $ESRMR_{U08}$ test successfully rejected the null hypothesis H_0: $SRMR_{pop} \geq EB_{upper}$ (where EB_{upper} = .08 and CI_{upper} = .017), indicating that the model fits the data such that the population SRMR value is likely to be negligible.

In this example, an SRMR equivalence test offered inferential information finding evidence of negligible misspecification beyond the descriptive information offered by the RMSEA, CFI, $SRMR_U$, and TLI estimates. Incidentally, the χ^2 goodness of fit test rejected the null hypothesis of perfect fit in this scenario. Thus, the descriptive fit measures along with the $ESRMR_{U08}$ test provide strong evidence of a well-fitting model without the need for ICTs or the logically problematic χ^2 goodness of fit test.

Discussion

When researchers evaluate model fit in SEM, they normally use various fit indices such as SRMR, RMSEA, and CFI. Unfortunately, instead of being interpreted as descriptive effect sizes, fit indices are often treated as ICTs, being compared to a cut-off value, with fit being viewed as acceptable or not (rather than interpreted along a continuum). The fit indices are usually presented alongside the χ^2 goodness of fit test, which has an implausible null hypothesis and has the logical flaw of non-significance being evidence in favour of the researcher's hypothesis that the model fits the data.

Equivalence testing in SEM allows proper population inference, beyond the χ^2 goodness of fit test, to be incorporated into model fit evaluation. Although various equivalence tests have been proposed for RMSEA and CFI (MacCallum et al., 1996; Yuan et al., 2016, Study 1), to the best of our knowledge no equivalence test based upon the SRMR has been proposed (despite it being one of the top three fit indices reported; Jackson et al., 2009). An equivalence test, presented alongside the SRMR index, provides formal inferential information (unlike an ICT), has a realistic null hypothesis (unlike the χ^2 goodness of fit test), and has an alternative hypothesis aligned with the researcher's goal of confirming adequate model fit (instead of rejecting exact fit).

This study proposed twelve potential variations of an SRMR-based equivalence test. The best-performing was the $ESRMR_{U08}$, with the $ESRMR_{U10A}$ a close second, but all twelve are openly available on OSF (https://osf.io/dq8sa/?view_only=df38766c227742d5b60e5d79b099df82) using the `neg.srmr()` function. Some of these variations were based on the original $SRMR_B$ sample statistic while others were based on the asymptotically unbiased estimate of SRMR proposed by Maydeu-Olivares (2017), $SRMR_U$. Some variations of the proposed tests included calculating the CI around SRMR in the method proposed by Maydeu-Olivares (which was only possible for tests based upon $SRMR_U$), while others incorporated a YHY bootstrap procedure to estimate the CI. Finally, for EBs, some of the proposed tests utilized EBs that were commonly used as cut-offs for SRMR ICTs (e.g., .05 and .08; Hu & Bentler, 1999). We termed these cut-offs unmodified EBs. We also used cut-offs that were more recently suggested by Shi et al. (2018) and Shi et al. (2022), which multiplied potential SRMR cut-off values (i.e., .05 or .10) by the average communality of the observed variables, $\overline{R}^2$. We referred to these more recent recommendations as modified EBs in our study.

Out of all the proposed equivalence tests in the study, the $ESRMR_{U08}$ and $ESRMR_{U10A}$ were the best-performing. The $ESRMR_{U08}$, or the Equivalence Based Fit Test for Unbiased SRMR with Equivalence Bound of .08, had slightly higher Type I error rates than some of the other equivalence tests when the misspecification was beyond the EB; however, rates were still quite low overall (.031 was the highest error rate, occurring at the smallest sample size of 50; rates were near zero, falling to .003 or less, by a sample size of 200). This test also had the best power rates of all equivalence tests at low sample sizes (e.g., $N = 50$) and comparable or better power rates than other equivalence tests at medium and large sample sizes (e.g., $N \geq 200$). We

did not initially evaluate the performance of this test at the EB (because we did not investigate an EB of .08 in the non-negligible misspecification at the EB condition); thus, as a follow up, we ran further simulations to confirm that the Type I error rate at the EB for $ESRMR_{U08}$ was indeed within Bradley's (1978) liberal criterion. The $ESRMR_{U10A}$, or the Equivalence Based Fit Test for Unbiased SRMR with Equivalence Bound of .10 Multiplied by Average Communality test, also had a good combination of Type I error rates and power; its power was comparable to the $ESRMR_{U08}$ at sample sizes of $N \geq 400$ and had reasonable Type I error control. Both tests used the Maydeu-Olivares (2017) method for estimating a CI around the unbiased SRMR; thus, whatever EB is chosen by the researcher, we can recommend it be compared to this specific CI type.

As with any equivalence test, choosing an EB is challenging. The EBs we selected for our simulation study, both modified and unmodified, were either ICT values commonly used by SEM researchers (e.g., .05 and .08; see Hu & Bentler, 1999) or values that resulted from newer methods for evaluating model misspecification using SRMR (e.g., $.05 \times \overline{R}^2$ and $.10 \times \overline{R}^2$; see Shi et al., 2018). Researchers should choose an EB that is most appropriate given the context of their study. Additional research has introduced dynamic fit indices derived through simulation that can be customizable ICTs for a given SEM model (McNeish & Wolf, 2021). These simulation-derived dynamic indices may be good candidates for EBs, but future research is necessary to understand how they would function as equivalence tests.

ICTs and the χ^2 goodness of fit test performed poorly in our study. ICTs did not exhibit any formal Type I error control when misspecification was at the EB. They tended to have higher power than their equivalence testing counterparts, which is expected because the SRMR index itself (rather than the upper bound of its CI) is compared to a cut-off value directly. However,

researchers using ICTs gain no interpretational benefits relative to equivalence tests. In other words, they do not offer any inferential information and cloud the use of descriptive fit indices, turning them into pseudo-hypothesis tests. As anticipated, when any model misspecification was present, the exact fit null hypothesis of the χ^2 goodness of fit test was rejected. In some conditions, this result occurred in tandem with equivalence tests providing evidence that certain kinds of misspecification were indeed "negligible". The use of this test is not recommended in the literature (Bentler & Bonett, 1980; Browne & Cudeck, 1992; Steiger, 2007; Hooper et al., 2008). Accordingly, we also do not recommend the χ^2 goodness of fit test and suggest equivalence tests be used to inferentially evaluate model fit instead.

Like all Monte Carlo simulation research, the findings in the present study are directly relevant only to the models and conditions we have compared. We anticipate, however, that the patterns across the methods compared will generalize to other sample sizes and model misspecifications. Further, more research is necessary to understand how the proposed equivalence tests perform when normality is violated, since the present study generated data from a multivariate normal distribution, or when alternative estimators are used.

This study introduced and evaluated the properties of inferential equivalence tests for SRMR that can be used and reported in tandem with descriptive fit indices. We hope that these tests will provide researchers with useful inferential tools for evaluating model fit in SEM.

CHAPTER FOUR

GENERAL DISCUSSION

Researchers investigating the fit of an SEM model normally rely on various fit indices and the χ^2 goodness of fit test. Often, fit indices are utilized as ICTs, being interpreted as pseudo-hypothesis tests, instead of effect sizes. To remedy this, equivalence testing has been introduced into model fit evaluation. The goal of equivalence testing, within this context, is to provide additional information for evaluating how well an SEM model fits the data, supplementing the reporting of fit indices and the χ^2 goodness of fit test. In practise, equivalence tests compare a given bound of an equivalence interval to the relevant bound of a CI. To date, equivalence tests have been proposed for RMSEA (MacCallum et al., 1996) and CFI (Yuan et al., 2016), but there has yet to be an equivalence test based upon SRMR. Although equivalence tests for CFI were already introduced (Yuan et al., 2016), their performance had yet to be evaluated alone or in tandem with RMSEA. These important research topics were explored in the two studies of the dissertation, which evaluated new and extant equivalence tests for fit indices in SEM. Each study introduced new equivalence tests for fit indices, conducted a Monte Carlo simulation, and introduced a real data example with code and equivalence testing functions that were made openly available.

Study 1 focused on equivalence tests for RMSEA and CFI, comparing previously introduced equivalence tests to the following newly proposed variations: EBF_{CFI}, $EBFB_{CFI}$, and $EBFB_{RMSEA}$. $EBFB_{CFI}$ and $EBFB_{RMSEA}$ are equivalence tests for CFI and RMSEA respectively, and both use a bootstrap procedure to calculate the relevant bound of a CI. The EBF_{CFI} is another equivalence test for CFI but uses an analytic computation of a CI bound derived by Yuan et al. (2016), instead of calculating the relevant CI bound via a bootstrap method. The EBF_{CFI} undoes

the previously-proposed Bonferroni-type correction by Yuan et al. to be more consistent with the equivalence testing literature, where a 100(1 -2α)% CI is standard. The $EBFB_{CFI}$ and $EBFB_{RMSEA}$, along with the original equivalence test for RMSEA (which we term EBF_{RMSEA}) first proposed by MacCallum et al. (1996), had the best combination of power and Type I error rates.

Study 2 proposed twelve potential options for an equivalence test base on the SRMR fit statistic. Various versions of these tests used differing methods for obtaining a CI (bootstrap vs. analytic), differing versions of the sample SRMR statistic (the original version proposed by Bentler, 1995, and the unbiased version proposed by Maydeu-Olivares, 2017), and different EBs (traditional values such as .05 or .08, and newer values that multiply .05 or .10 by the average communality of the observed indicators). The Monte Carlo study demonstrated that the $ESRMR_{U08}$, or the Equivalence Based Fit Test for Unbiased SRMR with EB of .08, had the best combination of power and Type I error rates.

In both studies, we compared numerous different types of tests using various EBs. In practice, when applying any equivalence test for fit indices, the specific EB selected will primarily depend upon the context of the research.

There is a slightly different treatment of modified EBs between the studies. Specifically, Study 1 does not assess the error rate at the modified EB for modified equivalence tests, evaluating the performance of modified RMSEA tests and CFI tests at EBs of .05 and .95, respectively. In contrast, Study 2 does assess Type I error control at a modified EB for a modified test (doing so for the tests that have an EB of .10 multiplied by the average communality of the indicators). In their conception, the modified EBs proposed by Yuan et al. (2016) that are evaluated in the first study were meant to be equivalence testing analogues of the

original cut-offs used for ICTs. For this reason, Study 1 asserts that it is reasonable to test their Type I error control at the original EB (i.e., either .05 or .95). In contrast, in Study 2, the modified EBs adapted from the ICTs proposed by Shi et al. (2018) and Shi et al. (2022) are not meant to be equivalence testing analogues of conventional ICT values of .05 or .08 (or any other value). Accordingly, due to their lack of relationship with any original EB, Study 2 does evaluate the Type I error control of modified SRMR equivalence tests at the modified EB.

Although the results of both Monte Carlo simulation studies are only directly relevant to the conditions investigated in both studies, it is anticipated that the pattern of results across the equivalence tests, ICTs, and the χ^2 goodness of fit test will generalize beyond the conditions investigated. Future research in this area may focus on equivalence tests for other fit indices such as TLI , focus on making robust versions of already-existing equivalence tests for fit indices, as well as explore the use of dynamic fit indices (see McNeish & Wolf, 2021) as EBs.

This dissertation has aimed to contribute to SEM literature by evaluating new and extant equivalence tests for fit indices for RMSEA, CFI, and SRMR. These procedures can be used in conjunction with the original fit indices (RMSEA, CFI, and SRMR) to gain a more complete understanding of model fit. Further, all equivalence tests evaluated in both studies can be easily conducted in R due to the openly available R functions `neg.rmsea()`, `neg.cfi()`, and `neg.srmr()`. It is hoped that the recommended tests from both studies will aid researchers in the assessment and interpretation of their own SEM models.

REFERENCES

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual.* Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Chalmers. R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248-280. doi:10.20982/tqmp.16.4.p248

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494. https://doi.org/10.1177/0049124108314720

Cheng, C. & Wu, H. (2017). Confidence intervals of fit indexes by inverting a bootstrap test. *Structural Equation Modeling, 24*(6), 870-880. https://doi.org/10.1080/10705511.2017.1333432

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology, 68*(2), 292-309. https://doi.org/10.1111/bmsp.12045

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research, 55*(2), 312-328. https://doi.org/10.1080/00273171.2019.1633617

Cribbie, R. A., Alter, U., Beribisky, N., Chalmers, R. P., Counsell, A., Farmus, L., & Martinez Gutierrez, N. (2023). negligible: A Collection of Functions for Negligible Effect/Equivalence Testing. R package version 0.1.3. https://CRAN.R-project.org/package=negligible

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*(3), 512–519. https://doi.org/10.1037/0033-2909.109.3.512

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Gomer, B., Jiang, G., & Yuan, K. H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(3), 371-389. https://doi.org/10.1080/10705511.2018.1545231

Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement, 61*(5), 741-758. https://doi.org/10.1177/00131640121971491

Hooper, D., Coughlan, J., & Mullen, M. (2008, June). Evaluating model fit: A synthesis of the structural equation modelling literature. In *7th European Conference on research methodology for business and management studies* (pp. 195-200).

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods, 14*(1), 6-23. https://doi.org/10.1037/a0014694

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333-351. https://doi.org/10.1207/S15328007SEM1003_1

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling, 26*(5), 757–777. https://doi.org/10.1080/10705511.2018.1562351

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*(3), 502–511. https://doi.org/10.1037/0033-2909.109.3.502

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

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: A Multidisciplinary Journal, 24*(1), 148-153. https://doi.org/10.1080/10705511.2016.1225260

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 indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2ß

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality, 37*(1), 48-75. https://doi.org/10.1016/S0092-6566(02)00534-2

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika, 82*(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7

McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on Lance, Beck, Fan, and Carter (2016*). Psychological Methods, 23*(1), 184–190. https://doi.org/10.1037/met0000157

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000425

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *The Journal of Experimental Education, 68*(3), 251-268. https://doi.org/10.1080/00220970009600095

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*(3), 353-377. https://doi.org/10.1207/S15328007SEM0803_2

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2020). simsem: SIMulated Structural Equation Modeling. R package version 0.5-15. https://CRAN.R-project.org/package=simsem

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17*(1), 1-14. https://doi.org/10.1037/a0026804

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. Retrieved from https://www.R-project.org/

Reifler, J., Scotto, T. J., & Clarke, H. D. (2011). Foreign policy beliefs in contemporary Britain: Structure and relevance. *International Studies Quarterly, 55*(1), 245-266. https://doi.org/10.1111/j.1468-2478.2010.00643.x

Reifler, J., Scotto, T. J., & Clarke, H. D. (2018). *Foreign Policy Beliefs in Contemporary Britain: Structure and relevance* (Harvard Dataverse; Version V1) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/TCJLUW

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(4), 369-379. https://doi.org/10.1080/10705519609540052

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research, 53*(5), 676-694. https://doi.org/10.1080/00273171.2018.1476221

Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research, 57*(2-3), 179-207. https://doi.org/10.1080/00273171.2020.1868965

Steiger, J. H. & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors* [Paper presentation]. IMPS 1980: Iowa City, IA, United States.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*(5), 893-898. https://doi.org/10.1016/j.paid.2006.09.017

Tanaka, J. S. (1987). " How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*(1), 134-146. https://doi.org/10.2307/1130296

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371-386. https://doi.org/10.1037/1082-989X.6.4.371

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society: Series B (Methodological), 37*(3), 406-412. https://doi.org/10.1111/j.2517-6161.1975.tb01554.x

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). CRC press.

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. https://doi.org/10.1080/10705511.2015.1065414

Yuan, K. H., Hayashi, K., & Yanagihara, H. (2007). A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research, 42*(2), 261-281. https://doi.org/10.1080/00273170701360662

Yuan, K. H., & Marshall, L. L. (2004). A new measure of misfit for covariance structure models. *Behaviormetrika, 31*(1), 67-90. https://doi.org/10.2333/bhmk.31.67

Yuan, K. H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80*(2), 379-405. https://doi.org/10.1007/s11336-013-9386-5

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 392-408. https://doi.org/10.1080/10705511.2015.1118692

TABLES

Table 1.
Names, confidence interval computation, equivalence bounds, and hypotheses for equivalence tests for RMSEA and CFI.

Test Name	Equivalence Based Fit Test Name	CI	Computation of CI	EB	Hypotheses
EBF_{RMSEA}	RMSEA; the not-close fit test for RMSEA by MacCallum et al. (1996)	100(1 – 2α) %	Iterative Algorithm	.05	H_0: $RMSEA_{pop} \geq$ MMES
$EBF_{RMSEA\text{-}A}$	RMSEA; using adjusted EB	100(1 – 2α) %	Iterative Algorithm	.05*	
$EBFB_{RMSEA}$	RMSEA; using YHY bootstrap for CI	100(1 – 2α) %	YHY Bootstrap	.05	H_1: $RMSEA_{pop} <$ MMES
$EBFB_{RMSEA\text{-}A}$	RMSEA; using adjusted EB and YHY bootstrap for CI	100(1 – 2α) %	YHY Bootstrap	.05*	
$EBF\alpha_{CFI\text{-}A}$	CFI; with modified CI using adjusted EB; by Yuan et al. (2016)	100(1 – α) %	Yuan method	.95*	
$EBF\alpha_{CFI}$	CFI; with modified CI	100(1 – α) %	Yuan method	.95	
EBF_{CFI}	CFI	100(1 – 2α) %	Yuan method	.95	H_0: $CFI_{pop} \leq$ MMES
$EBF_{CFI\text{-}A}$	CFI; using adjusted EB	100(1 – 2α) %	Yuan method	.95*	
$EBFB_{CFI}$	CFI; using YHY bootstrap for CI	100(1 – 2α) %	YHY Bootstrap	.95	H_1: $CFI_{pop} >$ MMES
$EBFB_{CFI\text{-}A}$	CFI; using adjusted EB and YHY bootstrap for CI	100(1 – 2α) %	YHY Bootstrap	.95*	

Note: CI = confidence interval; EB = equivalence bound; * = Yuan et al. modified equivalence bound that is analogous to .05 for RMSEA or .95 for CFI; MMES = minimally meaningful effect size or minimally tolerable amount of misspecification

Table 2.
Type I error rates for RMSEA and CFI equivalence tests and ICTs for Models 1, 2, and 3.

Model	N	Original Equivalence Bounds Tests					Modified Equivalence Bounds Tests					ICT	
		EBFB CFI	EBF CFI	EBFα CFI	EBFB RMSEA	EBF RMSEA	EBFB CFI-A	EBF CFI-A	EBFα CFI-A	EBFB RMSEA-A	EBF RMSEA-A	CFI	RMSEA
1	50	.000	.021	.006	.000	.016	.862	.638	.448	.046	.389	.467	.396
	75	.011	.017	.008	.000	.034	.877	.652	.480	.333	.438	.449	.443
	100	.049	.022	.012	.000	.041	.876	.690	.533	.435	.457	.502	.456
	200	.072	.026	.007	.007	.048	.874	.720	.570	.576	.537	.522	.528
	400	.071	.025	.014	.068	.044	.858	.703	.561	.580	.542	.508	.518
	800	.049	.020	.008	.076	.063	.837	.669	.523	.562	.517	.490	.547
	1000	.068	.029	.015	.067	.050	.816	.678	.538	.536	.495	.515	.496
	5000	.043	.024	.003	.067	.062	.244	.131	.070	.186	.180	.454	.523
2	50	.000	.005	.004	.000	.006	.040	.344	.209	.000	.177	.221	.174
	75	.000	.005	.001	.000	.019	.349	.475	.299	.002	.297	.318	.288
	100	.000	.012	.004	.000	.016	.527	.530	.371	.110	.396	.384	.390
	200	.020	.016	.004	.025	.036	.727	.629	.452	.352	.453	.442	.453
	400	.042	.017	.007	.046	.051	.758	.617	.474	.461	.483	.467	.502
	800	.038	.015	.004	.048	.047	.731	.615	.468	.435	.443	.465	.465
	1000	.054	.031	.013	.050	.054	.767	.635	.489	.479	.482	.477	.482
	5000	.044	.024	.011	.041	.041	.333	.233	.138	.453	.468	.442	.407
3	50	.000	.017	.010	.000	.007	.345	.885	.793	.000	.246	.288	.249
	75	.000	.046	.025	.000	.015	.638	.883	.836	.062	.372	.391	.371
	100	.007	.045	.029	.000	.019	.715	.867	.810	.206	.418	.408	.413
	200	.032	.066	.032	.028	.044	.826	.910	.858	.437	.462	.482	.460
	400	.050	.080	.044	.041	.037	.828	.887	.833	.514	.498	.490	.508
	800	.059	.085	.050	.047	.040	.807	.861	.801	.506	.479	.523	.502
	1000	.064	.093	.050	.064	.056	.793	.862	.785	.508	.476	.522	.501
	5000	.045	.079	.041	.052	.047	.216	.284	.211	.267	.261	.509	.495

Note: *N* = sample size; equivalence tests names are defined in Table 1; ICT = informal check test. Type I error rates for CFI and RMSEA come from different configurations of each model.

Table 3.
Proportion of indications of retaining perfect fit for χ^2 goodness of fit test, rejections of not close fit for RMSEA equivalence tests, rejections of not improved fit for CFI equivalence tests, and good fit for ICTs. Model 1.

N	Fit	Original Equivalence Bounds Tests					Modified Equivalence Bounds Tests					ICT		GoF
		EBFB CFI	EBF CFI	EBFα CFI	EBFB RMSEA	EBF RMSEA	EBFB CFI-A	EBF CFI-A	EBFα CFI-A	EBFB RMSEA-A	EBF RMSEA-A	CFI	RMSEA	
50	NN	.000	.011	.007	.000	.006	.837	.593	.417	.012	.124	.406	.126	.484
	Ne	.000	.105	.050	.000	.034	.992	.914	.824	.081	.474	.803	.475	.884
	P	.000	.174	.113	.000	.057	1.00	.955	.881	.117	.529	.852	.532	.880
75	NN	.012	.019	.006	.000	.000	.865	.643	.474	.061	.095	.440	.097	.328
	Ne	.164	.246	.138	.000	.060	.996	.969	.925	.478	.563	.922	.566	.856
	P	.250	.329	.198	.000	.081	.979	.990	.956	.575	.650	.963	.653	.912
100	NN	.047	.026	.006	.000	.001	.859	.655	.505	.069	.072	.433	.072	.197
	Ne	.514	.391	.234	.000	.057	1.00	.995	.975	.575	.599	.973	.599	.836
	P	.682	.551	.441	.000	.132	1.00	.999	.993	.731	.730	.989	.731	.906
200	NN	.044	.017	.010	.000	.000	.805	.585	.435	.010	.009	.381	.008	.008
	Ne	.927	.836	.728	.005	.105	1.00	1.00	1.00	.736	.704	1.00	.694	.719
	P	.991	.972	.924	.014	.326	1.00	1.00	1.00	.940	.913	1.00	.909	.926
400	NN	.028	.006	.002	.000	.000	.760	.566	.405	.000	.000	.347	.000	.000
	Ne	.999	.998	.992	.251	.184	1.00	1.00	1.00	.859	.832	1.00	.822	.442
	P	1.00	1.00	1.00	.834	.772	1.00	1.00	1.00	.999	.999	1.00	.999	.946
800	NN	.018	.005	.001	.000	.000	.658	.490	.337	.000	.000	.291	.000	.000
	Ne	1.00	1.00	1.00	.410	.348	1.00	1.00	1.00	.906	.892	1.00	.890	.099
	P	1.00	1.00	1.00	.994	.991	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.959
1000	NN	.015	.005	.001	.000	.000	.584	.418	.283	.000	.000	.257	.000	.000
	Ne	1.00	1.00	1.00	.467	.415	1.00	1.00	1.00	.948	.932	1.00	.933	.026
	P	1.00	1.00	1.00	1.00	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.941
5000	NN	.000	.000	.000	.000	.000	.009	.003	.000	.000	.000	.044	.000	.000
	Ne	1.00	1.00	1.00	.972	.997	1.00	1.00	1.00	.998	.998	1.00	1.00	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.935

Note: *N* = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined in Table 1; ICT = informal check test; GoF = χ^2 goodness of fit test.

Table 4.
Proportion of indications of retaining perfect fit for χ^2 goodness of fit test, rejections of not close fit for RMSEA equivalence tests, rejections of not improved fit for CFI equivalence tests, and good fit for ICTs. Model 2.

N	Fit	Original Equivalence Bounds Tests					Modified Equivalence Bounds Tests					ICT		GoF
		EBFB CFI	EBF CFI	EBFα CFI	EBFB RMSEA	EBF RMSEA	EBFB CFI-A	EBF CFI-A	EBFα CFI-A	EBFB RMSEA-A	EBF RMSEA-A	CFI	RMSEA	
50	NN	.000	.004	.003	.000	.003	.035	.287	.168	.000	.079	.165	.079	.232
	Ne	.000	.037	.022	.000	.023	.208	.651	.482	.000	.266	.473	.262	.551
	P	.000	.050	.018	.000	.020	.282	.782	.636	.000	.381	.625	.376	.687
75	NN	.000	.003	.000	.000	.000	.281	.395	.245	.001	.085	.237	.084	.140
	Ne	.000	.128	.067	.000	.056	.836	.910	.802	.009	.481	.801	.475	.646
	P	.126	.222	.137	.000	.099	.925	.967	.905	.019	.677	.912	.669	.882
100	NN	.000	.005	.001	.000	.000	.431	.434	.303	.010	.078	.296	.074	.082
	Ne	.044	.249	.146	.000	.083	.955	.955	.903	.268	.622	.906	.611	.642
	P	.989	.486	.337	.000	.214	.991	.990	.979	.502	.832	.978	.825	.847
200	NN	.009	.007	.001	.000	.000	.572	.464	.315	.008	.013	.283	.013	.000
	Ne	.783	.754	.605	.183	.261	1.00	.999	.995	.766	.853	.993	.852	.441
	P	1.00	.987	.967	.692	.810	1.00	1.00	1.00	.991	.994	1.00	.994	.912
400	NN	.011	.004	.002	.000	.000	.482	.362	.232	.001	.002	.217	.002	.000
	Ne	.998	.997	.984	.594	.619	1.00	1.00	1.00	.971	.976	1.00	.977	.100
	P	1.00	1.00	1.00	.982	.998	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.934
800	NN	.006	.003	.001	.000	.000	.364	.241	.150	.000	.000	.146	.000	.000
	Ne	1.00	1.00	1.00	.920	.920	1.00	1.00	1.00	.998	.997	1.00	.998	.002
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.949
1000	NN	.003	.001	.000	.000	.000	.336	.211	.121	.000	.000	.108	.000	.000
	Ne	1.00	1.00	1.00	.964	.964	1.00	1.00	1.00	.999	.999	1.00	.999	.001
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.949
5000	NN	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.001	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.935

Note: *N* = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined in Table 1; ICT = informal check test; GoF = χ^2 goodness of fit test.

Table 5.
Proportion of indications of retaining perfect fit for χ^2 goodness of fit test, rejections of not close fit for RMSEA equivalence tests, rejections of not improved fit for CFI equivalence tests, and good fit for ICTs. Model 3.

N	Fit	Original Equivalence Bounds Tests					Modified Equivalence Bounds Tests					ICT		GoF
		EBFB CFI	EBF CFI	EBFα CFI	EBFB RMSEA	EBF RMSEA	EBFB CFI-A	EBF CFI-A	EBFα CFI-A	EBFB RMSEA-A	EBF RMSEA-A	CFI	RMSEA	
50	NN	.000	.021	.011	.000	.003	.249	.822	.746	.000	.106	.222	.106	.360
	Ne	.000	.121	.073	.000	.028	.722	.979	.966	.000	.377	.650	.379	.741
	P	.000	.167	.100	.000	.033	.835	.989	.975	.000	.468	.712	.471	.779
75	NN	.000	.027	.009	.000	.001	.553	.814	.755	.012	.093	.301	.093	.248
	Ne	.000	.321	.240	.000	.076	.960	.996	.990	.116	.591	.874	.589	.805
	P	.003	.461	.345	.000	.094	.989	.999	.998	.181	.685	.924	.684	.879
100	NN	.001	.026	.012	.000	.001	.604	.844	.768	.032	.071	.310	.068	.136
	Ne	.177	.540	.422	.000	.108	.989	1.00	.998	.497	.686	.929	.683	.801
	P	.305	.678	.571	.000	.186	.999	1.00	.999	.631	.792	.976	.787	.890
200	NN	.015	.032	.014	.000	.001	.636	.769	.691	.016	.020	.300	.019	.003
	Ne	.915	.961	.923	.258	.324	1.00	1.00	1.00	.888	.911	1.00	.905	.748
	P	.991	.997	.991	.478	.572	1.00	1.00	1.00	.979	.984	1.00	.984	.936
400	NN	.007	.014	.009	.000	.000	.517	.643	.523	.000	.000	.224	.000	.000
	Ne	1.00	1.00	1.00	.745	.737	1.00	1.00	1.00	.987	.985	1.00	.987	.568
	P	1.00	1.00	1.00	.984	.983	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.941
800	NN	.001	.003	.001	.000	.000	.317	.406	.321	.000	.000	.112	.000	.000
	Ne	1.00	1.00	1.00	.986	.986	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.202
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.950
1000	NN	.000	.003	.000	.000	.000	.270	.360	.274	.000	.000	.076	.000	.000
	Ne	1.00	1.00	1.00	.996	.997	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.099
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.948
5000	NN	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000	.001	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1000	1.00	1.00	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1000	1.00	1.00	.947

Note: *N* = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined ‘in Table 1; ICT = informal check test; GoF = χ^2 goodness of fit test.

Table 6.
Names, confidence interval computation, equivalence bounds, and hypotheses for SRMR equivalence tests.

Test Name	Equivalence Based Fit Test Name	Computation of 100(1 – 2α) % CI	EB
$ESRMR_{U05}$	Unbiased SRMR with Equivalence Bound of .05	Maydeu-Olivares method	.05
$ESRMRY_{U05}$	Unbiased SRMR with Bootstrapping and Equivalence Bound of .05	YHY Bootstrap	.05
$ESRMRY_{05}$	Biased SRMR with Bootstrapping and Equivalence Bound of .05	YHY Bootstrap	.05
$ESRMR_{U08}$	Unbiased SRMR with Equivalence Bound of .08	Maydeu-Olivares method	.08
$ESRMRY_{U08}$	Unbiased SRMR with Bootstrapping and Equivalence Bound of .08	YHY Bootstrap	.08
$ESRMRY_{08}$	Biased SRMR with Bootstrapping and Equivalence Bound of .08	YHY Bootstrap	.08
$ESRMR_{U05A}$	Unbiased SRMR with Equivalence Bound of .05 Multiplied by Average Communality	Maydeu-Olivares method	$.05 \times \overline{R}^2$
$ESRMRY_{U05A}$	Unbiased SRMR with Bootstrapping and Equivalence Bound of .05 Multiplied by Average Communality	YHY Bootstrap	$.05 \times \overline{R}^2$
$ESRMRY_{05A}$	Biased SRMR with Bootstrapping and Equivalence Bound of .05 Multiplied by Average Communality	YHY Bootstrap	$.05 \times \overline{R}^2$
$ESRMR_{U10A}$	Unbiased SRMR with Equivalence Bound of .10 Multiplied by Average Communality	Maydeu-Olivares method	$.10 \times \overline{R}^2$
$ESRMRY_{U10A}$	Unbiased SRMR with Bootstrapping and Equivalence Bound of .10 Multiplied by Average Communality	YHY Bootstrap	$.10 \times \overline{R}^2$
$ESRMRY_{10A}$	Biased SRMR with Bootstrapping and Equivalence Bound of .10 Multiplied by Average Communality	YHY Bootstrap	$.10 \times \overline{R}^2$

Note: CI = confidence interval; EB = equivalence bound; $\overline{R}^2$ = average communality of observed variables

Table 7.
Proportion of lack of rejections of equivalence tests and good fit indications for $SRMR_B$ and $SRMR_U$ ICTs for non-negligible misspecification at the equivalence bound of .05.

Model	N	Original ESRMR Tests						Modified ESRMR Tests						ICTs	
		$_{U05}$	Y_{U05}	Y_{05}	$_{U08}$	Y_{U08}	Y_{08}	$_{U05A}$	Y_{U05A}	Y_{05A}	$_{U10A}$	Y_{U10A}	Y_{10A}	U-.05	B-.05
1	50	.067	.000	.000	.371	.263	.000	.000	.000	.000	.197	.008	.000	.571	.017
	75	.171	.008	.000	.407	.571	.034	.000	.000	.000	.229	.094	.000	.553	.053
	100	.172	.088	.000	.442	.708	.322	.002	.000	.000	.216	.234	.001	.534	.097
	200	.095	.209	.025	.609	.889	.742	.017	.002	.000	.166	.377	.143	.499	.199
	400	.052	.212	.069	.855	.975	.945	.002	.003	.000	.181	.463	.269	.524	.298
	800	.060	.176	.101	.990	.999	.999	.000	.000	.000	.241	.524	.394	.501	.347
	1000	.059	.177	.097	.995	1.00	1.00	.000	.000	.000	.291	.557	.453	.517	.392
	5000	.048	.099	.077	1.00	1.00	1.00	.000	.000	.000	.747	.865	.847	.494	.441
2	50	.168	.000	.000	.432	.329	.000	.000	.000	.000	.185	.005	.000	.530	.000
	75	.188	.046	.000	.556	.794	.002	.027	.000	.000	.211	.101	.000	.553	.000
	100	.133	.179	.000	.650	.953	.239	.038	.000	.000	.158	.247	.000	.571	.000
	200	.072	.433	.004	.840	.995	.949	.004	.002	.000	.105	.528	.032	.514	.030
	400	.053	.512	.104	.987	1.00	.999	.000	.002	.000	.102	.666	.212	.505	.110
	800	.043	.474	.198	1.00	1.00	1.00	.000	.000	.000	.109	.669	.369	.492	.197
	1000	.048	.476	.205	1.00	1.00	1.00	.000	.000	.000	.111	.717	.408	.499	.219
	5000	.038	.270	.178	1.00	1.00	1.00	.000	.000	.000	.282	.702	.612	.524	.395
3	50	.160	.001	.000	.418	.394	.000	.000	.000	.000	.192	.017	.000	.533	.000
	75	.146	.052	.000	.547	.791	.024	.018	.000	.000	.176	.156	.000	.548	.001
	100	.104	.162	.000	.635	.902	.352	.032	.000	.000	.137	.239	.000	.543	.009
	200	.058	.259	.006	.901	.994	.943	.002	.000	.000	.107	.406	.033	.524	.049
	400	.047	.288	.048	.995	1.00	1.00	.000	.000	.000	.127	.503	.158	.530	.139
	800	.033	.239	.071	1.00	1.00	1.00	.000	.000	.000	.133	.533	.269	.495	.210
	1000	.044	.258	.095	1.00	1.00	1.00	.000	.000	.000	.191	.594	.355	.521	.267
	5000	.036	.155	.080	1.00	1.00	1.00	.000	.000	.000	.545	.805	.722	.511	.401

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; ICT = informal check test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 8.
Proportion of lack of rejections of equivalence tests and good fit indications for $SRMR_B$ and $SRMR_U$ ICTs for non-negligible misspecification at the equivalence bound of .10 multiplied by the average communality of observed indicators.

Model	N	Original ESRMR Tests						Modified ESRMR Tests						ICTs	
		$_{U05}$	Y_{U05}	Y_{05}	$_{U08}$	Y_{U08}	Y_{08}	$_{U05A}$	Y_{U05A}	Y_{05A}	$_{U10A}$	Y_{U10A}	Y_{10A}	U-.10A	B-.10A
1	50	.046	.001	.000	.298	.234	.000	.000	.000	.000	.155	.009	.000	.580	.043
	75	.128	.013	.000	.301	.520	.037	.000	.000	.000	.160	.081	.000	.576	.102
	100	.094	.056	.000	.313	.590	.234	.002	.000	.000	.115	.161	.002	.539	.125
	200	.038	.127	.005	.479	.803	.647	.005	.000	.000	.076	.265	.065	.569	.258
	400	.006	.071	.022	.673	.917	.848	.000	.001	.000	.056	.205	.097	.507	.325
	800	.004	.030	.011	.904	.986	.969	.000	.000	.000	.043	.170	.110	.530	.382
	1000	.002	.023	.006	.960	.995	.991	.000	.000	.000	.033	.157	.093	.492	.373
	5000	.000	.000	.000	1.00	1.00	1.00	.000	.000	.000	.032	.081	.060	.482	.433
2	50	.186	.000	.000	.411	.319	.000	.000	.000	.000	.199	.003	.000	.565	.000
	75	.143	.034	.000	.504	.779	.005	.017	.000	.000	.158	.086	.000	.570	.001
	100	.113	.175	.000	.581	.908	.230	.034	.000	.000	.130	.244	.000	.558	.008
	200	.054	.343	.001	.793	.997	.934	.002	.001	.000	.075	.452	.016	.514	.041
	400	.021	.404	.056	.969	1.00	.998	.000	.001	.000	.047	.536	.120	.510	.122
	800	.012	.319	.092	1.00	1.00	1.00	.000	.000	.000	.033	.499	.227	.491	.238
	1000	.008	.306	.109	1.00	1.00	1.00	.000	.000	.000	.035	.495	.258	.510	.275
	5000	.001	.039	.018	1.00	1.00	1.00	.000	.000	.000	.039	.237	.167	.484	.367
3	50	.148	.000	.000	.363	.332	.000	.000	.000	.000	.157	.003	.000	.523	.001
	75	.103	.041	.000	.493	.733	.023	.004	.000	.000	.120	.102	.000	.577	.009
	100	.085	.123	.000	.578	.869	.289	.020	.000	.000	.112	.196	.000	.560	.017
	200	.036	.177	.002	.834	.990	.894	.000	.000	.000	.062	.285	.020	.528	.068
	400	.016	.148	.012	.973	.999	.993	.000	.000	.000	.038	.306	.057	.516	.165
	800	.004	.104	.014	1.00	1.00	1.00	.000	.000	.000	.035	.259	.105	.522	.250
	1000	.008	.071	.018	1.00	1.00	1.00	.000	.000	.000	.036	.251	.096	.521	.277
	5000	.000	.002	.001	1.00	1.00	1.00	.000	.000	.000	.033	.123	.070	.508	.404

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; ICT = informal check test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 9.
Proportion of retaining perfect fit for χ^2 goodness of fit test and rejections of equivalence tests. Model 1.

N	Fit	Original ESRMR Tests						Modified ESRMR Tests						GoF
		$_{U05}$	Y_{U05}	Y_{05}	$_{U08}$	Y_{U08}	Y_{08}	$_{U05A}$	Y_{U05A}	Y_{05A}	$_{U10A}$	Y_{U10A}	Y_{10A}	
50	NN	.003	.001	.000	.025	.033	.001	.000	.000	.000	.002	.001	.000	.025
	Ne	.369	.018	.000	.723	.635	.029	.003	.000	.000	.514	.163	.000	.830
	P	.381	.020	.000	.787	.697	.026	.006	.000	.000	.569	.194	.000	.899
75	NN	.001	.001	.000	.020	.025	.006	.001	.000	.000	.001	.001	.000	.002
	Ne	.555	.188	.000	.873	.927	.458	.050	.000	.000	.652	.508	.008	.856
	P	.595	.228	.000	.897	.949	.558	.047	.000	.000	.695	.594	.025	.909
100	NN	.000	.000	.000	.007	.010	.002	.000	.000	.000	.000	.000	.000	.000
	Ne	.591	.491	.000	.945	.982	.854	.175	.000	.000	.723	.728	.096	.826
	P	.709	.623	.001	.979	.998	.932	.182	.002	.000	.838	.861	.168	.920
200	NN	.000	.000	.000	.003	.005	.003	.000	.000	.000	.000	.000	.000	.000
	Ne	.811	.902	.519	.998	1.00	1.00	.334	.123	.000	.927	.970	.848	.730
	P	.932	.981	.791	1.00	1.00	1.00	.594	.222	.000	.983	.998	.974	.923
400	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.970	.990	.961	1.00	1.00	1.00	.403	.411	.013	.996	.999	.998	.460
	P	1.00	1.00	1.00	1.00	1.00	1.00	.829	.900	.142	1.00	1.00	1.00	.924
800	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.577	.672	.301	1.00	1.00	1.00	.092
	P	1.00	1.00	1.00	1.00	1.00	1.00	.991	1.00	.985	1.00	1.00	1.00	.941
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.636	.723	.416	1.00	1.00	1.00	.031
	P	1.00	1.00	1.00	1.00	1.00	1.00	.998	1.00	.998	1.00	1.00	1.00	.947
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.999	1.00	.997	1.00	1.00	1.00	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.951

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined in Table 1; GoF = χ^2 goodness of fit test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 10.
Proportions of good fit indications for $SRMR_B$ and $SRMR_U$ ICTs. Model 1.

N	Fit	$SRMR_B$ ICTs				$SRMR_U$ ICTs			
		.05	.08	.05A	.10A	.05	.08	.05A	.10A
50	NN	.000	.051	.000	.001	.038	.294	.003	.030
	Ne	.090	.790	.001	.284	.860	.987	.660	.920
	P	.147	.864	.003	.426	.901	.992	.746	.938
75	NN	.001	.050	.000	.001	.013	.205	.001	.014
	Ne	.306	.963	.001	.591	.921	.995	.697	.965
	P	.454	.982	.017	.706	.946	1.00	.782	.983
100	NN	.000	.042	.000	.000	.013	.166	.000	.002
	Ne	.520	.993	.008	.774	.950	.999	.698	.981
	P	.698	1.00	.040	.905	.988	1.00	.837	.997
200	NN	.000	.023	.000	.000	.000	.072	.000	.000
	Ne	.935	1.00	.114	.986	.992	1.00	.784	.998
	P	.988	1.00	.356	1.00	.999	1.00	.941	1.00
400	NN	.000	.008	.000	.000	.000	.019	.000	.000
	Ne	.999	1.00	.445	1.00	.999	1.00	.887	1.00
	P	1.00	1.00	.906	1.00	1.00	1.00	.997	1.00
800	NN	.000	.001	.000	.000	.000	.001	.000	.000
	Ne	1.00	1.00	.811	1.00	1.00	1.00	.965	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	.880	1.00	1.00	1.00	.973	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; ICT = informal check test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 11.
Proportion of retaining perfect fit for χ^2 goodness of fit test and rejections of equivalence tests. Model 2.

N	Fit	Original ESRMR Tests						Modified ESRMR Tests						GoF
		$_{U05}$	Y_{U05}	Y_{05}	$_{U08}$	Y_{U08}	Y_{08}	$_{U05A}$	Y_{U05A}	Y_{05A}	$_{U10A}$	Y_{U10A}	Y_{10A}	
50	NN	.002	.000	.000	.021	.026	.000	.000	.000	.000	.001	.000	.000	.013
	Ne	.524	.003	.000	.816	.611	.000	.052	.000	.000	.560	.038	.000	.611
	P	.594	.006	.000	.877	.681	.000	.058	.000	.000	.645	.057	.000	.701
75	NN	.000	.000	.000	.013	.033	.000	.000	.000	.000	.000	.000	.000	.002
	Ne	.597	.235	.000	.940	.972	.097	.307	.000	.000	.643	.380	.000	.733
	P	.727	.233	.000	.970	.983	.101	.383	.000	.000	.781	.432	.000	.814
100	NN	.000	.000	.000	.006	.025	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.717	.613	.000	.982	.998	.741	.409	.001	.000	.777	.734	.002	.780
	P	.845	.702	.000	.998	1.00	.847	.572	.000	.000	.882	.842	.000	.868
200	NN	.000	.000	.000	.000	.007	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.928	.983	.204	1.00	1.00	1.00	.373	.078	.000	.957	.991	.507	.707
	P	.987	1.00	.345	1.00	1.00	1.00	.678	.173	.000	.994	1.00	.726	.910
400	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.991	1.00	.992	1.00	1.00	1.00	.424	.508	.000	.996	1.00	.999	.492
	P	1.00	1.00	1.00	1.00	1.00	1.00	.895	.927	.000	1.00	1.00	1.00	.911
800	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.613	.850	.089	1.00	1.00	1.00	.124
	P	1.00	1.00	1.00	1.00	1.00	1.00	.993	1.00	.919	1.00	1.00	1.00	.950
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.670	.897	.223	1.00	1.00	1.00	.035
	P	1.00	1.00	1.00	1.00	1.00	1.00	.999	1.00	.994	1.00	1.00	1.00	.937
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.946

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined in Table 1; GoF = χ^2 goodness of fit test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 12.
Proportions of good fit indications for $SRMR_B$ and $SRMR_U$ ICTs. Model 2.

N	Fit	$SRMR_B$ ICTs				$SRMR_U$ ICTs			
		.05	.08	.05A	.10A	.05	.08	.05A	.10A
50	NN	.000	.003	.000	.000	.023	.233	.001	.017
	Ne	.000	.392	.000	.008	.862	.992	.611	.898
	P	.000	.459	.000	.011	.923	.997	.695	.942
75	NN	.000	.009	.000	.000	.007	.164	.000	.007
	Ne	.010	.861	.000	.084	.993	.998	.622	.954
	P	.021	.919	.000	.123	.967	.999	.768	.978
100	NN	.000	.007	.000	.000	.000	.118	.000	.000
	Ne	.085	.980	.000	.242	.967	1.00	.691	.975
	P	.154	.997	.000	.384	.992	1.00	.842	.997
200	NN	.000	.002	.000	.000	.000	.042	.000	.000
	Ne	.839	1.00	.000	.905	.998	1.00	.786	.999
	P	.961	1.00	.001	.984	1.00	1.00	.946	1.00
400	NN	.000	.001	.000	.000	.000	.002	.000	.000
	Ne	.997	1.00	.022	.999	1.00	1.00	.851	1.00
	P	1.00	1.00	.376	1.00	1.00	1.00	.996	1.00
800	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	.403	1.00	1.00	1.00	.963	1.00
	P	1.00	1.00	.989	1.00	1.00	1.00	1.00	1.00
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	.565	1.00	1.00	1.00	.979	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; ICT = informal check test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 13.
Proportion of retaining perfect fit for χ^2 goodness of fit test and rejections of equivalence tests. Model 3.

N	Fit	Original ESRMR Tests						Modified ESRMR Tests						GoF
		$_{U05}$	Y_{U05}	Y_{05}	$_{U08}$	Y_{U08}	Y_{08}	$_{U05A}$	Y_{U05A}	Y_{05A}	$_{U10A}$	Y_{U10A}	Y_{10A}	
50	NN	.005	.000	.000	.031	.018	.000	.000	.000	.000	.004	.000	.000	.015
	Ne	.524	.005	.000	.840	.756	.001	.040	.000	.000	.593	.114	.000	.744
	P	.610	.021	.000	.889	.856	.003	.065	.000	.000	.690	.204	.000	.802
75	NN	.000	.000	.000	.025	.026	.000	.000	.000	.000	.000	.000	.000	.001
	Ne	.621	.312	.000	.969	.988	.377	.276	.000	.000	.712	.553	.001	.807
	P	.731	.481	.000	.989	.996	.603	.420	.000	.000	.820	.717	.003	.867
100	NN	.000	.000	.000	.011	.018	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.724	.662	.000	.995	1.00	.912	.395	.001	.000	.840	.828	.010	.806
	P	.856	.836	.000	.999	1.00	.980	.587	.003	.000	.914	.935	.066	.914
200	NN	.000	.000	.000	.001	.003	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.955	.991	.543	1.00	1.00	1.00	.448	.214	.000	.986	.998	.864	.781
	P	.998	1.00	.855	1.00	1.00	1.00	.714	.433	.000	1.00	1.00	.976	.918
400	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	.999	1.00	.998	1.00	1.00	1.00	.545	.639	.000	1.00	1.00	1.00	.574
	P	1.00	1.00	1.00	1.00	1.00	1.00	.952	.984	.007	1.00	1.00	1.00	.949
800	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.756	.926	.395	1.00	1.00	1.00	.176
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.996	1.00	1.00	1.00	.954
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	.863	.961	.594	1.00	1.00	1.00	.100
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.954
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00.	.000
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.954

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; equivalence tests names are defined in Table 1; GoF = χ^2 goodness of fit test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

Table 14.
Proportions of good fit indications for $SRMR_B$ and $SRMR_U$ ICTs. Model 3.

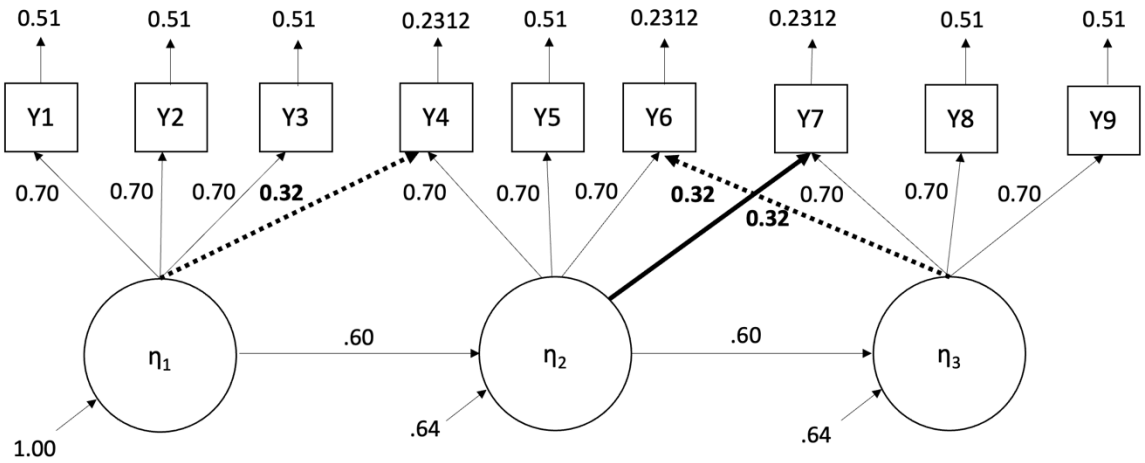
N	Fit	$SRMR_B$ ICTs				$SRMR_U$ ICTs			
		.05	.08	.05A	.10A	.05	.08	.05A	.10A
50	NN	.000	.007	.000	.000	.044	.297	.007	.042
	Ne	.005	.647	.000	.061	.888	.998	.645	.939
	P	.011	.783	.000	.119	.932	.998	.727	.964
75	NN	.000	.016	.000	.000	.017	.235	.000	.013
	Ne	.058	.964	.000	.263	.956	1.00	.691	.981
	P	.157	.994	.000	.445	.984	1.00	.796	.997
100	NN	.000	.016	.000	.000	.002	.198	.000	.002
	Ne	.255	.997	.000	.561	.982	1.00	.733	.993
	P	.493	1.00	.001	.757	.995	1.00	.858	.999
200	NN	.000	.014	.000	.000	.000	.087	.000	.000
	Ne	.944	1.00	.010	.987	1.00	1.00	.857	1.00
	P	.999	1.00	.067	1.00	1.00	1.00	.973	1.00
400	NN	.000	.003	.000	.000	.000	.027	.000	.000
	Ne	1.00	1.00	.223	1.00	1.00	1.00	.930	1.00
	P	1.00	1.00	.800	1.00	1.00	1.00	.999	1.00
800	NN	.000	.001	.000	.000	.000	.003	.000	.000
	Ne	1.00	1.00	.765	1.00	1.00	1.00	.981	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	.913	1.00	1.00	1.00	1.00	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	NN	.000	.000	.000	.000	.000	.000	.000	.000
	Ne	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: N = sample size; P = perfect fit; Ne = negligible misspecification; NN = non-negligible misspecification outside of equivalence bound; ICT = informal check test; A after test name = Shi et al. (2018) and Shi et al. (2022) modification of multiplying either .05 or .10 by the average communality of the observed indicators.

FIGURES

Figure 1

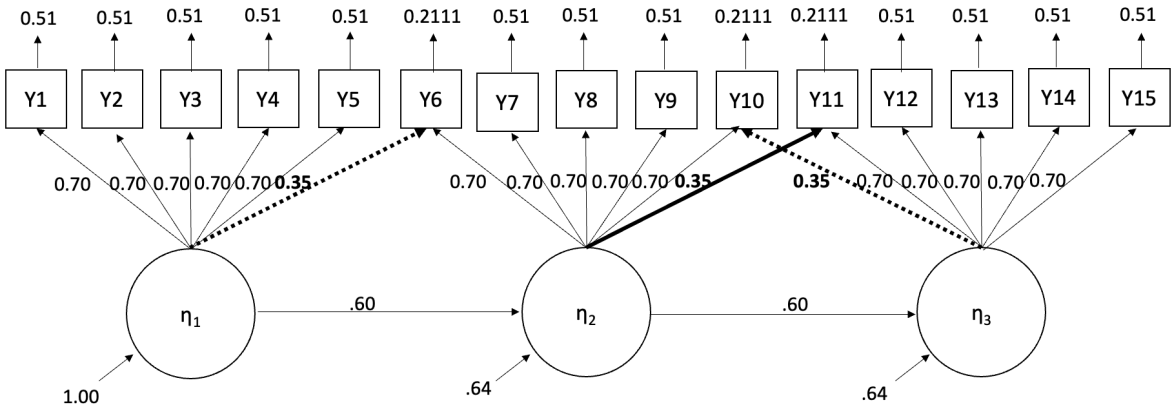
Population Generating Model 1 for RMSEA and CFI Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Dotted paths, in addition to bolded paths, are omitted in non-negligible specifications. Bolded loadings correspond to bolded/dotted paths for clarity of view. In the negligible misspecification condition, population RMSEA and CFI were 0.0390 and 0.9916, respectively. In the non-negligible misspecification (outside of the equivalence bound) condition, RMSEA and CFI were 0.0963 and 0.9445, respectively.

Figure 2

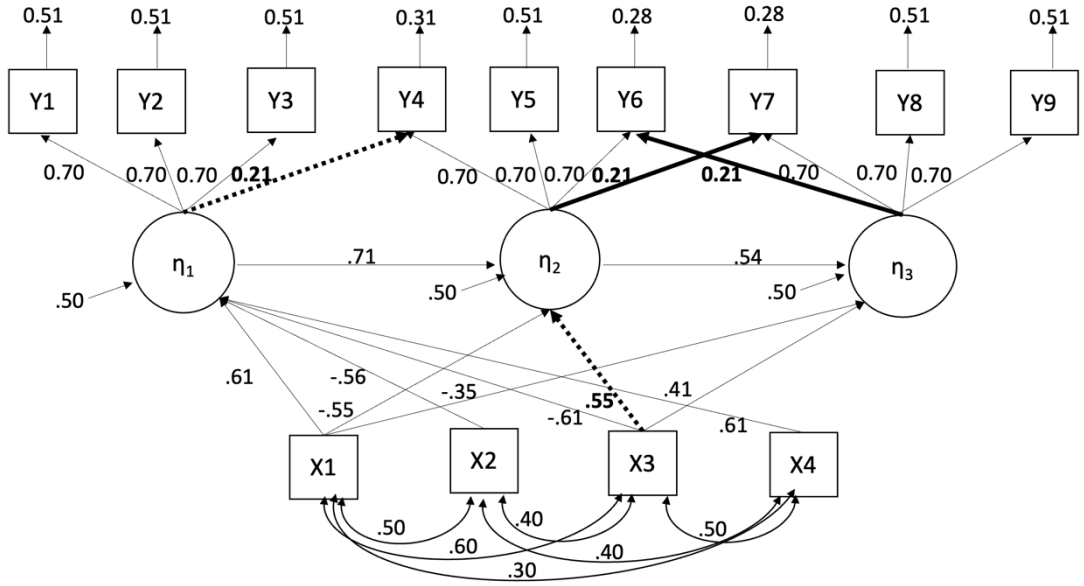
Population Generating Model 2 for RMSEA and CFI Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Dotted paths, in addition to bolded paths, are omitted in non-negligible specifications. Bolded loadings correspond to bolded/dotted paths for clarity of view. In the negligible misspecification condition, population RMSEA and CFI were 0.0360 and 0.9854, respectively. In the non-negligible misspecification (outside of the equivalence bound) condition, RMSEA and CFI were 0.0706 and 0.9425, respectively.

Figure 3

Population Generating Model 3 for RMSEA and CFI Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Dotted paths, in addition to bolded paths, are omitted in non-negligible specifications. Bolded loadings correspond to bolded/dotted paths for clarity of view. In the negligible misspecification condition, population RMSEA and CFI were 0.0272 and 0.9930, respectively. In the non-negligible misspecification (outside of the equivalence bound) condition, RMSEA and CFI were 0.0770 and 0.9412, respectively.

Figure 4

Monte Carlo Simulation Conditions for RMSEA and CFI Equivalence Tests

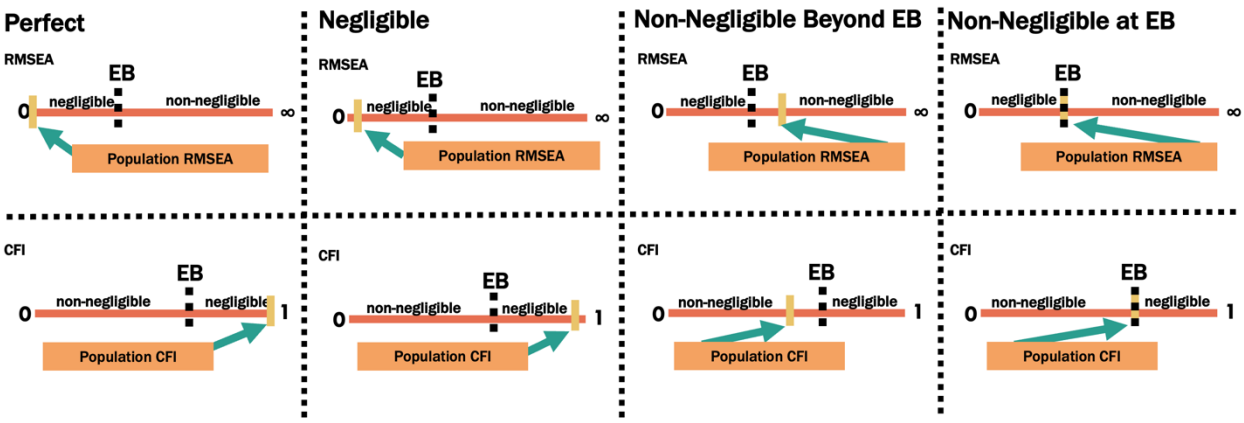


Figure 5

Proposed CFA Model Structure for the Humor Styles Questionnaire

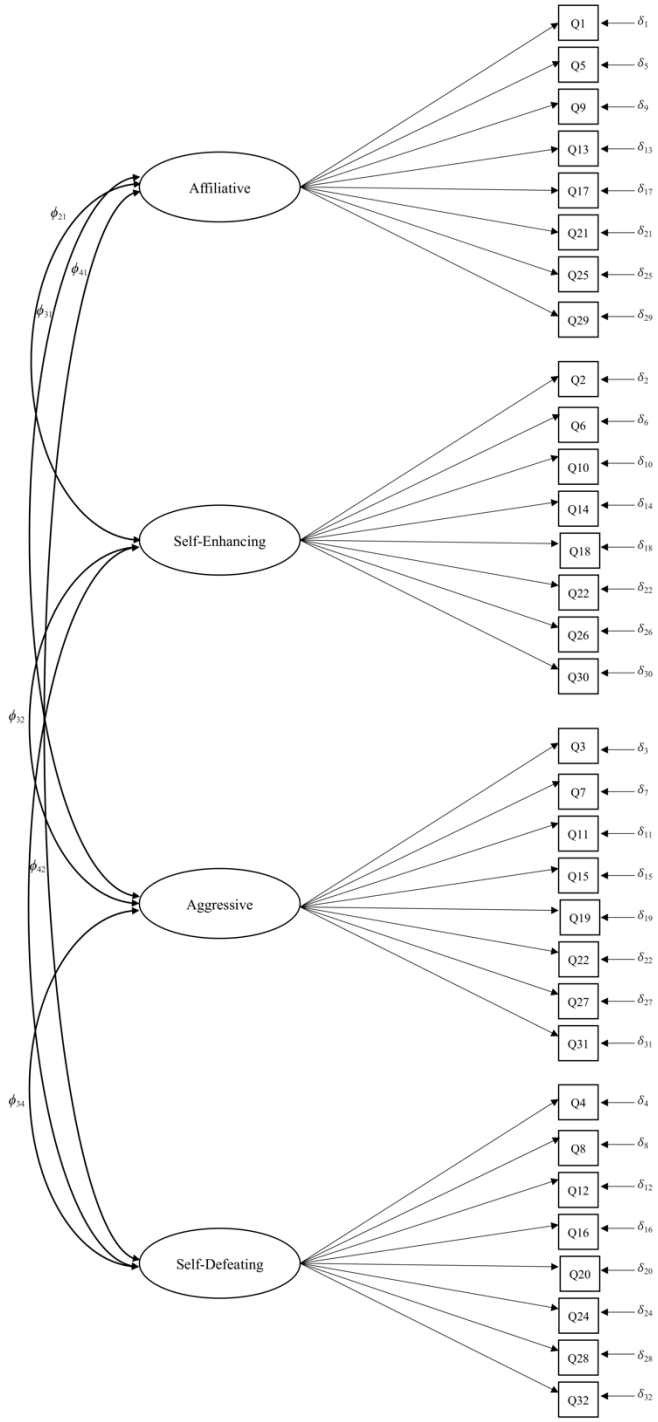
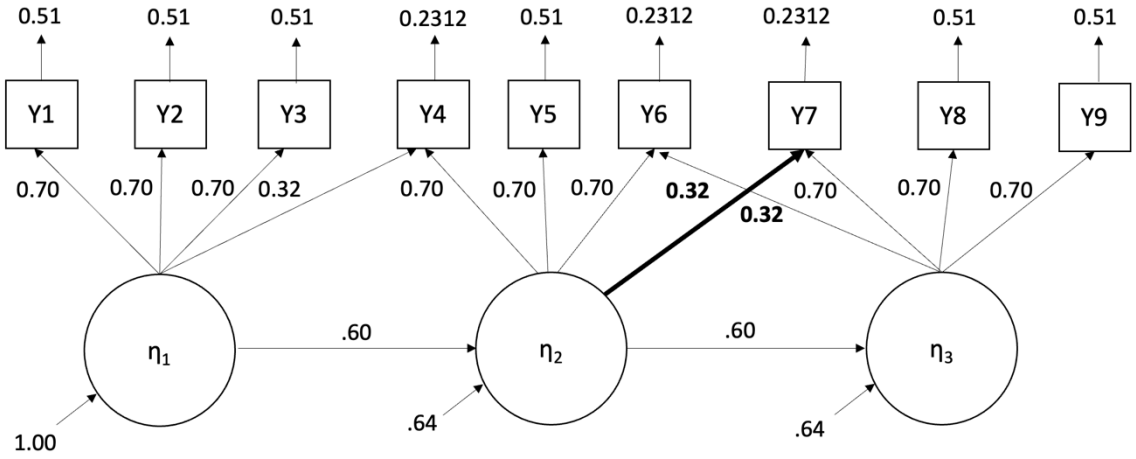


Figure 6

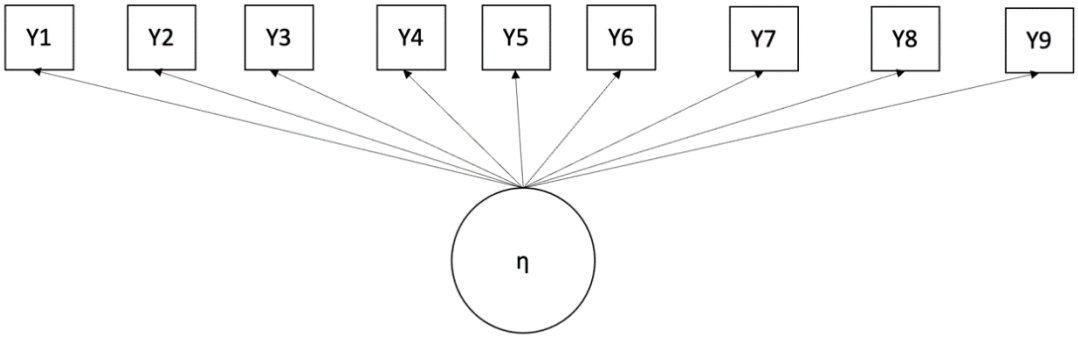
Population Generating Model 1 for SRMR Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Bolded loadings correspond to bolded paths for clarity of view. Population SRMR is 0.01993201 for the negligible misspecification condition.

Figure 7

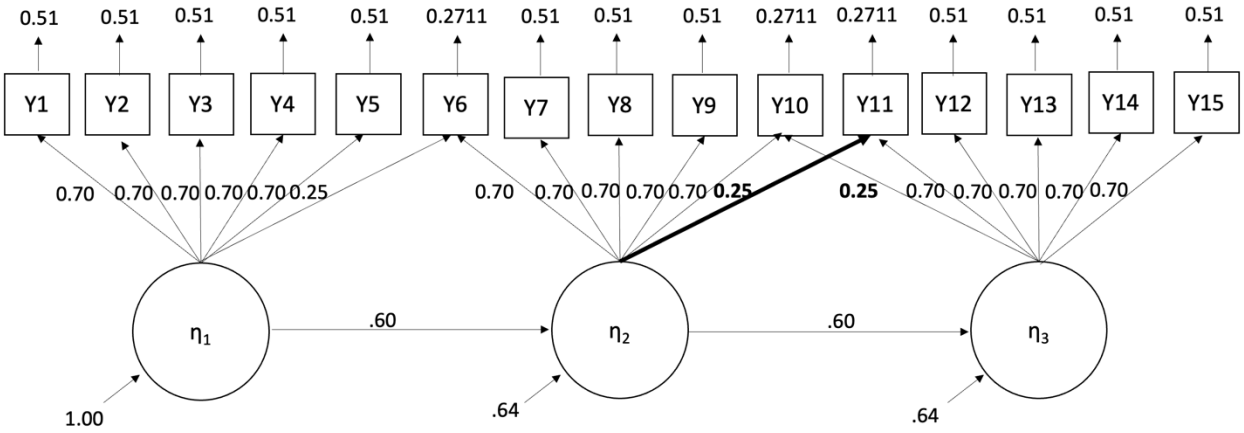
Non-Negligible Misspecification outside of the Equivalence Bound for Model 1



Note. Loadings are standardized. Population SRMR is 0.09879457.

Figure 8

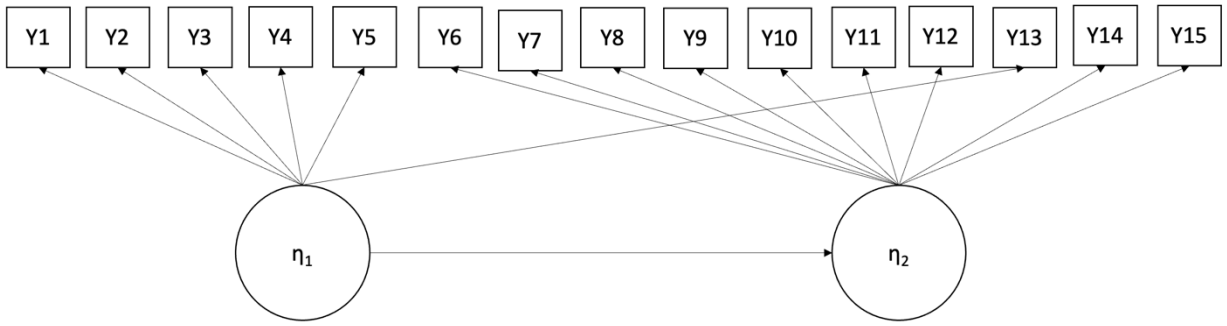
Population Generating Model 2 for SRMR Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Bolded loadings correspond to bolded paths for clarity of view. Population SRMR is 0.01928947 for the negligible misspecification condition.

Figure 9

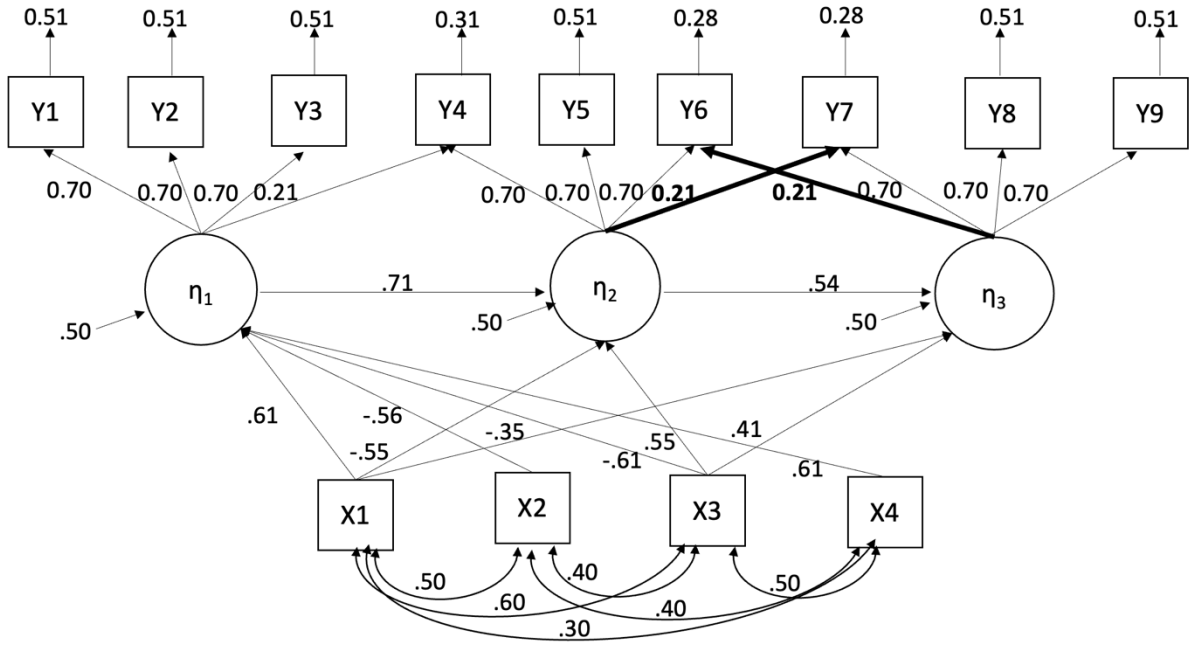
Non-Negligible Misspecification outside of the Equivalence Bound for Model 2



Note. Loadings are standardized. Population SRMR is 0.1013469.

Figure 10

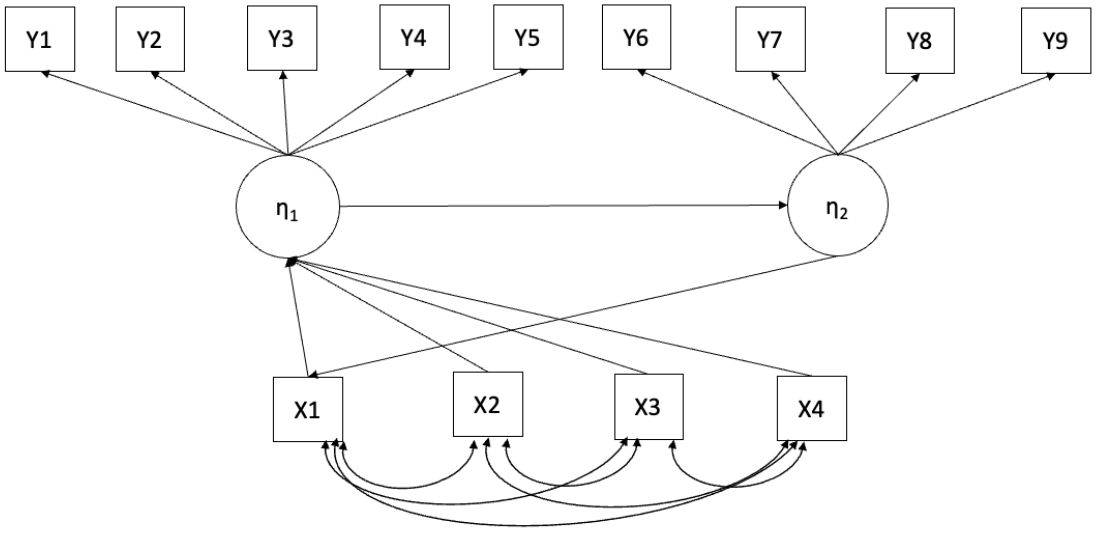
Population Generating Model 3 for SRMR Equivalence Tests: Adapted from Chen et al. (2008)



Note. Loadings are standardized. Bolded paths are omitted in negligible specifications. Bolded loadings correspond to bolded paths for clarity of view. Population SRMR is 0.01743751 for the negligible misspecification condition.

Figure 11

Non-Negligible Misspecification outside of the Equivalence Bound for Model 3



Note. Loadings are standardized. Population SRMR is 0.09852503.

Figure 12

Monte Carlo Simulation Conditions for SRMR Equivalence Tests

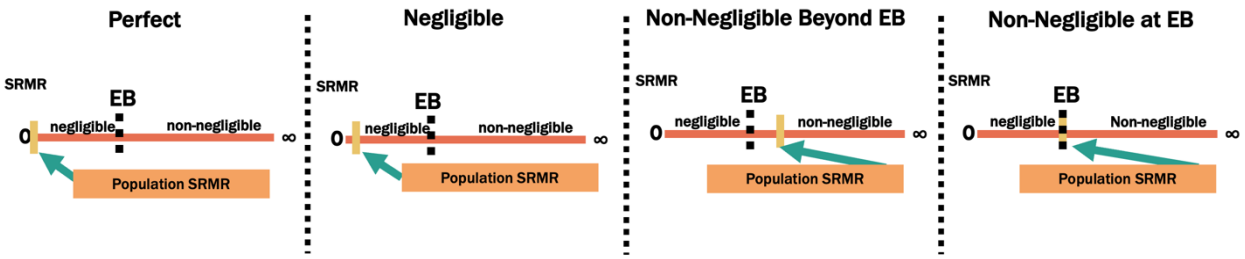


Figure 13

Two Factor Confirmatory Factor Analysis of British Foreign Policy from Reifler et al. (2011)