4-26-2019

Misfit at the Intersection of Measurement Quality and Model Size: A Monte Carlo Examination of Methods for Detecting Structural Model Misspecification

Graham Rifenbark
University of Connecticut - Storrs, graham.rifenbark@uconn.edu

Misfit at the Intersection of Measurement Quality and Model Size: A Monte Carlo Examination of Methods for Detecting Structural Model Misspecification

Graham Gilbert Rifenbark, PhD

University of Connecticut, 2019

The evaluation of misfit in structural equation models (SEM) is an area of great importance, as the structural parameters are those we utilize to make population level inferences about some causal process. Recently, structural fit indices (SFIs) have been advanced due to the influence of the measurement model on the approximate fit indices (AFIs). First, AFIs are overly weighted by the measurement model (McDonald & Ho, 2002). Second, AFI cut-offs were not determined in the context of varying measurement quality; as a result, model fit appears to improve as measurement quality decreases, known as the *reliability paradox* (Hancock & Mueller, 2011). The approach advanced by Hancock and Mueller (2011) requires two stages of estimation, whereas, the approach advanced by Lance, Beck, Fan, and Carter (2016) is accomplished by simultaneous estimation of all model parameters. The focus of this dissertation was to understand the sampling distribution of the various SFIs. This was accomplished in the Type I Error simulation and was used to empirically select cut-off-values. The secondary focus was to examine the relative performance of the SFIs in their ability to detect a misspecified mean structure, covariance structure, and simultaneous misspecifications. This study was executed in the context of multiple group models where the misspecifications were in the form of true differences between populations. Central to the study across simulations was the impact construct reliability had on Type I errors and power rates for the SFIs, as well as how they performed relative to AFIs. Major findings were as follows. The two-stage approach of Hancock and Mueller (2011) should not be utilized, as it was found to be impacted by measure reliability. The structural measures of fit outperformed the global measures of fit regardless of the type of misfit (e.g., mean or covariance). Measures of fit were more sensitive to measure reliability when the covariance structure was misspecified, compared with when the mean structure was misspecified.

Misfit at the Intersection of Measurement Quality and Model Size: A Monte Carlo Examination of Methods for Detecting Structural Model Misspecification

Graham Gilbert Rifenbark

B.A., University of Kansas, **2010**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

Misfit at the Intersection of Measurement Quality and Model Size: A Monte Carlo Examination of Methods for Detecting Structural Model Misspecification

Presented by

Graham Gilbert Rifenbark, B.A.

Co-Major Advisor _____
D. Betsy McCoach

Co-Major Advisor _____
H. Jane Rogers

Associate Advisor _____
Hariharan Swaminathan

Associate Advisor _____
Eric Loken

Associate Advisor _____
Christopher Rhoads

University of Connecticut

2019

ACKNOWLEDGMENTS

Many colleagues, mentors, and friends have helped and encouraged me along the way. Dr. Sarah Bunnell taught me introductory statistics at the University of Kansas (KU) and urged me dive head first into statistics. Dr. Little hired me as a Research Assistant at the Center for Research Methods and Data Analysis (CRMDA). I had the pleasure of collaborating with Drs. Todd Little, Kris Preacher, Carol Woods, Wei Wu, Pascal Deboeck, and Paul Johnson. It was my time at the CRMDA that convinced me to earn a Ph.D. with a concentration on quantitative methodology. I am grateful for all of my experiences and opportunities that were made available to me at the CRMDA. Not to mention, all of the great friends and collaborators I made there.

I acknowledge the Graduate Research Assistantship support I received over the years from faculty in the Department of Educational Psychology and all of the opportunities that I had in this role. I am grateful for all of the support I received from the faculty in the Research Methods, Measurement, and Evaluation (RMME) program. The RMME faculty challenged me to think critically about research and as a result, I learned a great deal from them. I would like to thank my committee members: Drs. Hariharan Swaminathan, Eric Loken, and Christopher Rhoads. Each of you gave me valuable feedback on this dissertation study and were willing and able to assist me along the way. I specifically want to thank my Co-Major Advisors: Drs. D. Betsy McCoach and H. Jane Rogers. This dissertation simply would not have been possible without your tireless support, of which I am entirely grateful. I would like to acknowledge the Booth Engineering Center for Advanced Technology High Performance Computing cluster which allowed me to conduct this dissertation study efficiently.

Finally, I would like to thank my family for all of their support over the years. My parents gave my brother and me an unforgettable childhood living overseas with rich experiences. I will always cherish these memories. That being said, I have immeasurable gratitude for all of the love and support my mother and father have given me over the years. This dissertation is for you.

Contents

List of Figures

List of Tables

1. Introduction

All models are misspecified to some degree, as the true model in the population is unknown. We approximate the true model to the best of our abilities and collect evidence that the model fits the data to a reasonable degree. When the model concerns relationships between unobservable phenomena, a latent variable (LV) approach is typically taken. LV models have both a measurement model and a structural model; however, model fit pertains to the model as a whole. Because the structural model and its parameters serve as the basis for making inferences to some population, evaluation of structural model fit is of utmost importance.

LV models posit that common variance among observed items [manifest variables (MVs)] is explained by the LV, which represents the measured construct. Alternatively, observed variables can be modeled as outcomes that are regressed onto the latent variable. The measurement model specifies the relationship between the latent and manifest variables. The structural model specifies the relationship between the measured latent constructs. Using the factor structure and the latent structure, a model-implied variance-covariance matrix and mean vector are derived. The evidence we collect on how well a given model approximates the true model is quantified as the distance between the observed and model-implied moments. This discrepancy is known as F and estimation is designed to minimize $\hat{F}_{ML}$. The F_{ML} value for a saturated model (assuming a population variance-covariance matrix) must be zero; therefore, $\hat{F}_{ML}$ can be interpreted as the distance from exact fit.

Assuming multivariate normality, a test statistic can be constructed that is χ^2 distributed to test the statistical significance of the deviation from exact fit. However, such a test of model fit is unrealistic (Box & Draper, 1987); therefore, a host of *approximate fit indices* (AFIs) have been proposed that *describe* model fit (Moshagen & Erdfelder, 2016). A shortcoming of AFIs is that their sampling distribution is unknown; therefore, it is impossible to determine conclusively what constitutes a poor or an acceptable fitting model based on their values, even though the majority of the AFIs depends on $\hat{F}_{ML}$. As a result, we must rely on cut-off values to be determined

empirically (via simulation) to determine what constitutes acceptable model fit.
The decision to retain or dismiss a LV model is made in large part based on the simulation work of Hu and Bentler (1999). The cut-offs proposed for common AFIs have become *golden thresholds* and ultimately determine the fate of a model. Unfortunately, it is common for these recommended cut-offs to be misused in practice because of differences between candidate models and the models used in the simulation (Ding, Velicer, & Harlow, 1995; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011), even though Hu and Bentler (1998) warn researchers not to over-generalize their findings. A key shortcoming of the Hu and Bentler (1998) study was their failure to investigate varying levels of construct reliability (Heene et al., 2011; Kenny, Kaniskan, & McCoach, 2015; McNeish, An, & Hancock, 2018), which Gagne and Hancock (2006) describe as a function of measurement quality and the number of indicators per factor (p:f).
In structural equation models (SEM) measurement quality is defined by the magnitude of standardized factor loadings. These parameters are standardized regression coefficients that indicate the strength of the relationship between the observed variables and the latent variable. It follows then, that factor loadings are interpreted as the rate at which an indicator can differentiate between those who are high and low on the measured construct. When the magnitude of standardized factor loadings is large, measurement quality is high, and as a result, so is construct reliability (Gagne & Hancock, 2006).
Misspecification of the structural model results in a greater distance between the observed and model-implied moments when measurement quality is high, compared with when measurement quality is low. This is especially true for elements in the model-implied variance-covariance matrix that correspond to MVs that load onto separate LVs (Moshagen & Auerswald, 2017). Therefore, models with high quality measures appear to be penalized when comparing the AFIs from their hypothesized model with the cut-offs proposed by Hu and Bentler (1999). Ultimately, the decision to reject a model due to its being an unreasonable approximation to the true model may be a consequence of both measurement quality and when to reject (i.e., cut-off values).

This phenomenon is known as the reliability paradox (Hancock & Mueller, 2011). Controlling for measurement quality, AFIs are differentially influenced by the size of the model. As the p:f ratio increases, the observed variance-covariance matrix is of a larger dimension, which in turn disadvantages some AFIs and will indicate unacceptable fit, while other AFIs are advantaged and indicate acceptable fit (Ding et al., 1995; Kenny et al., 2015). Some AFIs are more sensitive to misspecification in the structural model than the measurement model and vice-versa (Hu & Bentler, 1999), and this behavior relates to whether or not the AFI is based on F_{ML}, which is weighted disproportionately by the measurement model (Moshagen & Auerswald, 2017).

Two approaches have been proposed to deal with these problems, and each offers a solution to evaluating the structural model without interference from the measurement model. The solution offered by Lance et al. (2016) evaluates structural model fit via latent variables, hereafter referred to as *SM-LV*, and was offered in response to the measurement parameters overwhelming the structural parameters. The solution given by Hancock and Mueller (2011) evaluates structural model fit via manifest variables, hereafter referred to as *SM-MV*, as a response to the reliability paradox. *Structural fit indices* (SFIs) result from the SM-LV and SM-LV approaches and are used to evaluate structural model fit. Similar to AFIs, the sampling distribution of SFIs is unknown and it is unclear what constitutes acceptable fit. The SM-MV approach is unique as it requires two stages of estimation: (1) a LV model is estimated, followed by (2) a path model which is estimated using information from Stage-1 (the LV model). On the other hand, the SM-LV results from single-stage estimation and, therefore, uncertainty of parameter estimates is not lost.

Research has demonstrated that a key challenge of two-stage methods is the manner in which uncertainty around parameter estimates from Stage-1 are incorporated (or taken into account) in the Stage-2 model (Levy, 2017). On its face, the SM-MV approach proposed by Hancock and Mueller (2011), there is no attempt made to incorporate the uncertainty from the initial model in the path model. For this reason, it is odd that McNeish and Hancock (2018) question the use of the Lance et al. (2016) approach

(SM-LV) and advocate the use of the untested SM-MV approach.

The goal of this dissertation was to evaluate the performance of the SM-MV and to the extent to which the two-stage nature of this approach impacts its behavior, and whether their was an advantage to single-stage estimation inherent in the SM-LV approach. This goal was evaluated in the context of detecting structural misfit in multiple group LV models. Both approaches can be applied in this context, yet their merits are unknown. The focus of this study was to understand how construct reliability affects the SM-MV and SM-LV approaches in their ability to detect structural model misfit. To this end, I conducted two Monte Carlo simulations in the context of modeling group differences, which commonly incorporates both the variance-covariance and mean structure. Ultimately, this study is meant to inform researchers who utilize this common LV model, enabling them to make informed decisions regarding the fit of their structural model. In turn, this will lead to robust inferences about the phenomena under investigation. Specifically, I examined the sampling distribution of the SFIs from the SM-MV and SM-LV approaches to understand their sensitivity to model and data characteristics. This was accomplished in the Type I Error simulation. I add to the literature by investigating the merits of SM-MV and SM-LV in three novel ways. Specifically, I examined the relative performance of the approaches for detecting varying levels of misspecifications in: (1) the mean structure, (2) the covariance structure, and (3) both the mean and covariance structures using a multiple group model. This was accomplished in the Power simulation, in which I utilized conventional approaches to detecting group differences, such as $\triangle\chi^2$, to determine the relative merits of the SM-LV and SM-MV approaches as these are unknown.

2. Literature Review

Latent Variable Models

To make inferences to some population about a phenomenon of interest require (1) specifying a system of equations, (2) choosing an estimation method [e.g., maximum likelihood (ML)], and (3) employing an estimation technique [e.g., expectation-maximization (EM)]. The estimation technique is tasked with minimizing some function and in LV models it is the *discrepancy fit function* (F_{ML}). F_{ML} serves as the basis from which structural equation models are judged with respect to their goodness-of-fit.

System of Equations. The equations presented below are the means from which hypotheses about free and fixed paths are tested. Confirmatory factor analysis (CFA) models differ from SEMs in that no causal structure among the latent variables is specified and therefore, all latent variables are exogenous. Free parameters are hypothesized to have nonzero estimates in the population. The set of parameter estimates that minimize the distance between the observed variance-covariance matrix and mean vector and their respective model-implied moments are *maximum likelihood estimates* because they maximize the likelihood of the data.

CFA. For this model, all LVs are treated as exogenous and their relations are freely estimated. $\hat{\Sigma}$ corresponds to the model-implied variance-covariance matrix, whereas, $\hat{\mu}$ corresponds to the model-implied mean vector.

$$\begin{aligned} y &= \tau + \Lambda x + e \\ Var(x) &= \Phi \\ Var(e) &= \Theta \\ \hat{\mu} &= \tau + \Lambda\alpha \\ \hat{\Sigma} = Var(y) &= \Lambda\Phi\Lambda^t + \Theta \end{aligned} \tag{1}$$

In Equation 1, Λ, the vector of factor loadings, represents unstandardized regression coefficients and corresponds to an indicator's ability to discriminate between those who are high and low on the latent construct; when Λ is standardized, this corresponds to

measurement quality. Θ is the error variance-covariance matrix and is usually assumed to be diagonal in factor analysis; these elements result from regressing the MVs on to the LV (i.e., observed variance that is unrelated to the latent construct). Φ is a symmetric matrix that represents the latent variance-covariance matrix; the diagonal elements are the variances of the construct in the population; and the off-diagonal elements are the covariances between LVs. With respect to the means, α, the vector of latent means provides an anchoring point for the construct(s) in the population. Finally, τ, the vector of manifest intercepts, correspond to the expected value of the observed variable when the latent construct equals zero.

SEM. When causal relationships are specified among LVs, $\hat{\Sigma}$ and $\hat{\mu}$ depend on parameters that stem from regressing certain LVs on other LVs. The linear model which gives rise to $\hat{\Sigma}$ and $\hat{\mu}$ is written as:

$$\mathbf{Y} = \tau + \mathbf{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}(\alpha + \zeta) + \epsilon, \tag{2}$$

$\mathbf{I}$ is an identity matrix with as many rows and columns as there are LVs. Latent regression coefficients are stored in a square matrix, $\mathbf{B}$, which has the same dimensions as $\mathbf{I}$. Equation 2 illustrates that the diagonals of $\mathbf{B}$ must be zero, while the free elements below the diagonal correspond to latent regression parameters (e.g., B[1, 2] is free when LV2 is regressed onto LV1). Equation 2 is at the individual level, where α is a vector of latent intercepts with the same dimensions as $\mathbf{B}$ and $\mathbf{I}$; ζ is a column vector that represents the individuals' deviations from the mean on each LV; and ϵ is a column vector with as many rows as there are MVs and corresponds to the individuals unique scores on each MV. From this linear model, the model-implied moments are derived as follows (Widaman & Thompson, 2003):

$$\begin{aligned}\hat{\mathbf{\Sigma}} = [(\hat{\tau} + \hat{\mathbf{\Lambda}}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha})(\hat{\tau} + \hat{\mathbf{\Lambda}}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha})'] + \\ [\hat{\mathbf{\Lambda}}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\mathbf{\Psi}}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Theta}}]\end{aligned} \tag{3}$$

$$\hat{\mu} = \hat{\tau} + \hat{\mathbf{\Lambda}}(\mathbf{I} - \hat{\mathbf{B}})^{-1}\hat{\alpha}. \tag{4}$$

With respect to the structural model, $\mathbf{\Psi}$ is a matrix with the same dimensions as $\mathbf{B}$

that corresponds to covariances between the latent disturbances. With respect to the measurement model, τ, $\mathbf{\Lambda}$, and $\mathbf{\Theta}$ are the same as in Equation 1.

One-Stage Versus Two-Stage Estimation. When single-stage estimation is employed, measurement and structural parameters from Equation 3 are simultaneously estimated. The benefit of utilizing single-stage estimation is that measurement error is explicitly modeled and in turn, the statistical significance of the resulting ML estimates can be appropriately assessed. This is due to the uncertainty of each parameter being accounted for via standard errors. Two-stage estimation approaches are defined as those that require fitting models in a serial fashion, whereby information from the initial model is used in a subsequent model. An example of a two-stage model is as follows. Consider a single LV that is estimated from Equation 1. Its factor scores (Φ) can be extracted and used to predict some observed outcome variable in a linear regression. It is reasonable to believe that the resulting factor scores are measured without error, as this is the purpose of LV models; however, the uncertainty of the factor scores that result from the initial LV model cannot be incorporated into the subsequent regression model.

An alternative approach is to incorporate the observed outcome variable into the latent variable model. In this way, the benefits of single-stage estimation can be realized. Although this appears to be promising, Levy (2017) points to a key problem with this approach. Because the measurement and structural parameters are estimated simultaneously, it is impossible for the LV to distinguish between observed variables that are indicators of it and those that are outcomes (Levy, 2017). As a result, the meaning of the LV then changes based on the observed outcome that is inserted into the model; this is known as interpretational confounding (Burt, 1976). To combat this problem, procedures have been advanced that require fixing measurement parameters to values that result from the CFA model [e.g., Tucker (1971)] and appropriately defines the LVs, however, the uncertainty around the measurement parameter estimates is not incorporated in the subsequent model.

Scaling and Identification. Due to the simultaneous estimation of parameters that pertains to both the observed and unobserved variables, the concept of scaling is important. Specifically, we can scale with respect to the observed variables and fix a single factor loading (λ_{11}) per latent variable to unity and its respective manifest intercept to zero (τ_{11}). By inserting these model constraints, we are employing the marker variable approach, and, as a result, the interpretation of Φ and α are in the metric of the observed variables (Bollen, 1989). Alternatively, model constraints can be placed on the latent variances (diagonal of Φ) and means (α), where the former is set to unity and the latter is fixed to zero and scaling is done with respect to the LVs (Bollen, 1989). As a consequence, the off-diagonals of Φ are interpreted as latent correlations and the parameter estimates from Λ and τ remain in their observed scale (Bollen, 1989). Not only do the above mentioned model constraints perform the important role of scaling the model parameters, but also their use is necessary in order to ensure parameter estimates are unique.

Information and degrees of freedom. Information (or knowns) stems from both the sample variance-covariance matrix and the mean vector. An inherent problem with LV models is the simultaneous presence of free parameters that must be estimated and correspond to either observed or unobserved variables. As such, there may not be enough information to guarantee estimates are unique, unless some form of model identification is employed (e.g., marker variable or fixed factor). For example, consider a CFA model that has simple structure (i.e., each MV loads onto only one LV) that contains a total of P observed variables - MVs; and a total of K unobserved variables - LVs.

$$INFO = \frac{P(P+1)}{2} + P = \frac{P*(P+3)}{2} \tag{5}$$

For concreteness, P = 25 and K = 5 (oblique factors); therefore, each LV contains 5 indicators. In such a situation we can estimate the total amount of mean and variance information available using Equation 5, and we find there are 350 unique pieces of information. In total, 25 belong to the mean structure and 325 belong to the covariance

structure. The degrees of freedom (df) for the covariance structure is determined as the number of free parameters in Λ, Θ, and Φ subtracted from 325. In a similar fashion, the df in the mean structure is determined as the number of free parameters in τ and α subtracted from 25.

With respect to the covariance structure, there are 25 free parameters in Λ, 25 in Θ, and 15 in Φ ($\frac{K*(K+1)}{2}$), totaling 65 free parameters. On the other hand, the mean structure possesses 25 free parameters in τ and 5 free parameters in α, totaling 30 free parameters. As a result, the degrees of freedom in the covariance structure is 325 - 65 or 260 and for the mean structure its degrees of freedom is 25 - 30 or -5. Even though there is enough information to estimate all of the covariance parameters, their uniqueness is not guaranteed. On the other hand, there is not enough information to estimate all 30 free parameters in the mean structure, rendering it under-identified. The remedy for these problems is to employ some form of scaling. As a result, 5 degrees of freedom are gained in both the covariance and mean structures, rendering the mean structure just-identified: 25 - 25 = 0 degrees of freedom; this changes the degrees of freedom in the covariance structure to 265: 325 - 60. In sum, this measurement model contains 85 free parameters and 265 degrees of freedom (350 - 85). Note that all of the global degrees of freedom stem from the covariance structure and none from the mean structure. As a result, the mean structure does not contribute to model misfit, as it is saturated (i.e., contains zero degrees of freedom making it just identified).

Basic Model Fit. Regardless of whether a CFA or a SEM model is estimated, convergence is achieved once F_{ML} has been minimized. Modeling the mean and covariance structure simultaneously, Browne and Arminger (1995) write the *discrepancy fit function* as:

$$F_{ML}(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}; \hat{\mu}, \mu) = (\mu - \hat{\mu})^t \hat{\mathbf{\Sigma}}^{-1}(\mu - \hat{\mu}) + \ln \left|\mathbf{\Sigma}\right| + tr(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}) - \ln \left|\hat{\mathbf{\Sigma}}\right| - q, \qquad (6)$$

where q corresponds to the number of freely estimated parameters across both the measurement and the structural models. Using $\hat{F}_{ML}$, a test statistic known as the likelihood test statistic (or T) is constructed as: $T = F_{ML} * N$. Assuming the observed data are multivariate normally distributed, T is distributed as χ^2 with degrees of

freedom equal to: $df = \frac{P(P+3)}{2} - q$ (Bollen, 1989). If the observed moments are perfectly reproduced by the model, exact fit is established. Given exact fit, both $\hat{F}_{ML}$ and $\hat{T}$ must equal zero.

Fit statistics versus indices. A criticism of fit statistics (e.g., T or χ^2) is that they are highly impacted by sample size. Specifically, statistical power to detect negligible model misfit using the χ^2 test statistic increases with sample size. It has been argued that tests of exact fit are unreasonable and therefore, *fit indices* have been proposed as another way to evaluate misfit (Moshagen & Erdfelder, 2016). Aside from this difference, fit statistics and fit indices differ in an important way. Specifically, the sampling distribution is known for the χ^2 test statistic; therefore, the significance of any departure from exact fit can be quantified. On the other hand, the sampling distributions of most fit indices (AFIs or GFIs) are unknown; therefore, quantifying model misfit becomes challenging. However, fit statistics and indices share two attractive attributes. First, both evaluate the fit of the entire model simultaneously (i.e., measurement and structural models are assessed jointly). Second, both are able to assess model fit without the need for an alternative hypothesized model to be estimated, unlike information criteria.

Global Fit Measures

In applied research, it is common for the *chi-square test statistic* or likelihood test statistic [T; Bollen (1989)], the *root mean squared error of approximation* [RMSEA; (Steiger & Lind, May, 1980)], the *McDonald's Measure of Centrality* [Mc; McDonald (1989)], the *standardized root mean square residual* [SRMR; Jöreskog and Sörbom (1981)], the *Tucker-Lewis index* [TLI; Tucker and Lewis (1973)], and the *comparative fit index* [CFI; Bentler (1990)] to be used. Each of these fit indices is formally presented in Chapter 3.

Classifying Fit Indices. The simplest classification of AFIs was given by Yuan (2005), who proposed a dichotomy where fit is evaluated either using test statistics (i.e., a function of F_{ML} and degrees of freedom) or residuals (i.e., deviations in the mean

vector and covariance matrix). SRMR is the only example of an index that is commonly reported that is residual-based.

Sun (2005) proposed a more thorough taxonomy for classifying AFIs. This taxonomy considers three dimensions, and further classifies indices on two finite attributes. The three dimensions are: (1) population or sample based, (2) absolute or incremental, and (3) includes adjustment for model complexity or not.

Population versus sample. The distinction between population and sample based indices lies with respect to the moments on which the discrepancy is based. *Population based indices* represent the discrepancy between the population and the model-implied covariance matrices ($\Sigma - \hat{\Sigma}$), and the population and model-implied mean vectors ($\mu - \hat{\mu}$). Because the population moments are unavailable, a sample estimate of the non-centrality parameter (λ) of the discrepancy is: $\lambda = T - df$. Due to the effect of sample size on T (i.e., $F_{ML} * N$), λ can be rescaled by $\frac{\lambda}{N-1}$. When the non-centrality parameter is rescaled in this manner, McDonald's d (McDonald, 1989) results. *Sample based indices* utilize sample moments instead and make the assumption that $S \sim \Sigma$ and $M \sim \mu$ approximate their respective moments in the population; therefore, discrepancy is quantified by $(S - \hat{\Sigma})$ and $(M - \hat{\mu})$, and $T = \hat{F} * N$ suffices.

Absolute versus incremental. The fit of a LV model falls somewhere between the worst fitting model (i.e., the independent null model) and the best fitting model (i.e., the saturated model). *Absolute fit indices* are evaluated with respect to the saturated model, which is a model that is just-identified (df = 0). *Incremental fit indices* are evaluated with respect to some baseline model, which most commonly is represented by the *independent null model*; however, there have been many other baseline models suggested for different modeling contexts (Little, Preacher, Selig, & Card, 2007; Widaman & Thompson, 2003; Wu & West, 2010). A difficulty encountered when specifying the baseline model relates to the mean structure and how it is specified, as it is not always clear. The baseline model is characterized to be an over-identified (i.e., few free parameters) model and in the case of the independent null model, observed variables are modeled such that they are orthogonal to one another, and only the mean

and variance are freely estimated for each MV, df > 0.

Adjustment for model complexity. Some indices make no attempt to adjust for the complexity of the model; however, some will do so in one of three ways utilizing the model's degrees of freedom (df): (1) some linear combination (i.e., information criteria); (2) dividing T by df; or (3) by estimating a parsimony index which is a function of the df from the hypothesized model (df_H) and the baseline model (df_B) (e.g., $PI = \frac{df_H}{df_B}$) and multiplying an incremental fit index by it (Williams & Holahan, 1994).

Reliability Paradox. Hancock and Mueller (2011) illustrated the reliability paradox through a population level SEM analysis with 3 exogenous and endogenous LVs. The population model was severely misspecified by omitting 4 regression paths and a correlated latent residual. They showed that given the same set of misspecifications, the RMSEA and SRMR worsened as measurement quality increased, as indicated by Λ; whereas, the incremental AFIs also generally decreased (indicating worse fit), with the exception of CFI, which indicated better fit at the high end of Λ. Further, they showed that the power of modification indices to detect the misspecified structural paths decreased as Λ decreased. Therefore, it is difficult to identify structural misspecification via modification indices when measurement quality is low. Aside from affecting the power of modification indices and model fit, measurement quality also was shown to affect the sampling variability of structural parameter estimates; for example, when $\Phi = 0.5$, standard error estimates were 0.107 and 0.037 for Λ equal to 0.40 and 0.95, respectively. Ultimately, these findings motivated Hancock and Mueller (2011) to propose their two-stage manifest variable approach (SM-MV) for evaluating structural model fit in isolation.

In a similar vein, research conducted by Heene et al. (2011) showed that the cut-offs Hu and Bentler (1999) suggested for AFIs are affected by measurement quality. The goal of the simulation conducted by Heene et al. (2011) was to highlight the effect of different population values of Θ, as a result of varying levels of measurement quality, on the ability of AFIs to detect the same structural misspecification when using Hu and Bentler (1999) cut-off values. The simulation was based on a three-factor solution and

investigated the effect model complexity had on AFIs (simple or complex, via cross-loadings); sample size (N = 150, 250, 500, 1000, and 2500); measurement quality (Low = $0.31 \leq \Lambda \leq 0.47$, Medium = $0.51 \leq \Lambda \leq 0.67$, High = $0.71 \leq \Lambda \leq 0.87$); and the number of items (15 or 45). For all simulation conditions, Φ was: 0.3, 0.4, or 0.5 in the population; however, these paths were fixed to zero in the estimation model. Therefore, by fixing a covariance that is nonzero in the population to zero $\hat{F}_{ML}$ must become larger as measurement quality increases (Moshagen & Auerswald, 2017). Using tracing rules, we can illustrate the mechanism from which the reliability paradox operates and understand Heene et al. (2011) findings.

Reliability paradox mechanism. As an example, take a two-factor model where $\Phi_{1,2} = 0.4$ in the population, and assume simple structure for Λ and that the measures are highly reliable (e.g., elements of $\Lambda = 0.8$). The population covariances for items from the same LV will be $0.8^2 = 0.64$, and the covariances among items from different factors will be $(0.8 * 0.4 * 0.8)$ or 0.256. If we place a misspecification into the structural model in a similar fashion as Heene et al. (2011), $\Phi_{1,2}$ is fixed to 0.0. As a result, the model-implied covariance among items from the same factor remains the same at 0.64, whereas the model-implied covariance between items from different factors is (0.8*0.0*0.8) or 0. This leads to a difference of 0.256 for each of the elements in $\mathbf{S} - \hat{\mathbf{\Sigma}}$ that load onto different LVs. Alternatively, if measurement quality is low (e.g., elements of $\Lambda = 0.3$), the covariances among common items are 0.3^2 or 0.09, and the population covariances among uncommon items are (0.3*0.4*0.3) or 0.036. After inserting the same structural misspecification ($\Phi_{1,2} = 0.0$), the model-implied covariances among common items remain 0.09; however, the model-implied covariance between items from two separate LVs is (0.3*0.0*0.3) or 0.0. Therefore, the difference estimates between $\mathbf{\Sigma}$ and $\hat{\Sigma}$ for items that load onto separate factors is 0.256 when $\Lambda = 0.8$ and 0.036 when $\Lambda = 0.3$. It follows then that the effect of measurement quality is absorbed into the estimation of $\hat{F}_{ML}$, which in turn affects T and the AFIs that depend on $\hat{F}_{ML}$ (Moshagen & Auerswald, 2017).

Moshagen and Auerswald (2017) showed this behavior via simulations where either the

F_{ML} or SRMR (which is independent of F_{ML}) remained constant across the population and the misspecified model. Moshagen and Auerswald (2017) accomplished this by estimating cross-loadings to account for the latent covariance being fixed to zero, ensuring that the same F_{ML} or SRMR estimate was obtained. In doing so, Moshagen and Auerswald (2017) illustrated that under these conditions the reliability paradox vanishes and as a consequence, T and RMSEA were no longer affected by varying levels of measurement quality, whereas the CFI remains affected by measurement quality due to its effect on the independent null model. For example when measurement quality is poor, the independent null model will provide a better fit to the data and thus create a ceiling effect on the CFI and TLI; when measurement quality is high the independent model will fit worse and therefore, CFI and TLI will not be impacted by a ceiling effect.

Structural Fit Measures

James, Mulaik, and Brett (1982) proposed two conditions to determine whether or not causal relationships have been correctly specified: *condition nine* (C9) and *condition ten* (C10). Recall that moving from a CFA model to a SEM model requires certain paths between LVs to be uni-directional and others to be zero. Therefore, C9 corresponds to a hypothesized non-zero relationship between LVs that is confirmed to be non-zero in the population, while, C10 corresponds to a hypothesized null relationship between LVs, that in fact is null in the population. I define the hypothesized causal model as the *target structural model* (SM_{target}), which imposes uni-directional paths between LVs and sets certain pathways to zero. The available approaches for assessing structural model fit can be classified as *conventional*, *latent*, or *two-stage* and are detailed below.

Conventional Measures. The conventional methods used to assess structural fit evaluate misfit from the absolute perspective. These approaches are the $\triangle\chi^2_{[\triangle df]}$ and the *root mean square error of approximation - path* [RMSEA-P; McDonald and Ho (2002)]. The former offers a test of exact fit, while the latter offers a test of close fit; however, information from the same two models is used to estimate them. The models used are the SM_{target} and its CFA model (or SM_{sat}).

Chi-square difference tests ($\triangle\chi^2_{[\triangle df]}$). Anderson and Gerbing (1988, 1992) show that C9 and C10 assumptions can be examined by omitting or inserting a nonzero path in a candidate model and carrying out sequential $\triangle\chi^2$ tests. Because the structural model is saturated in the CFA model, the χ^2 and degrees of freedom that result offer the best fitting structural model. Therefore, when paths are fixed to 0 in the hypothesized target model (SM_{target}), a statistical test can be performed to evaluate the effect on model fit.

RMSEA-P. McDonald and Ho (2002) proposed the structural analog for RMSEA. This was motivated by a systematic review of SEM studies, which confirmed two common occurrences: the proportion of degrees of freedom belong to the measurement model and the adequacy of structural model fit stemming from GFIs do not assess the structural model in isolation. Using RMSEA-P, McDonald and Ho (2002) showed that in many instances, poor structural fit is masked by the widely reported global RMSEA. To estimate RMSEA-P, the $\triangle T$ and $\triangle df$ between SM_{sat} and SM_{target} are used to estimate d_p $[\frac{\triangle T - \triangle df}{\triangle df}]$.

Latent Measures - SM-LV. Using C9 and C10 as a basis, Lance et al. (2016) proposed three types of indices to investigate structural model fit: (1) T, which does not take model complexity into account, (2) the ratio of T to degrees of freedom - $\frac{T}{df}$, and (3) the non-centrality parameter - $\lambda = T - df$. Regardless of the measure (C9 or C10) or the method from which they are estimated, all rely on the estimation of three LV models: the null (SM_{null}), the target (SM_{target} - same as the conventional methods), and the saturated (SM_{sat} - same as the conventional methods). The interpretation of these SFIs differs from the traditional sense of absolute and incremental evaluation of fit. Specifically, the C9 indices are not determined based only on the null model; likewise, the C10 indices are not based solely on the saturated model. Instead, they are formed based on the distance between the null and the saturated models. As such, the denominator remains the same for C9 and C10 indices, with the only difference emerging in the numerator of their equations (See Equation 7).

$$C9 = \frac{T_{SM_{null}} - T_{SM_{target}}}{T_{SM_{null}} - T_{SM_{sat}}};$$
$$C10 = \frac{T_{SM_{target}} - T_{SM_{sat}}}{T_{SM_{null}} - T_{SM_{sat}}}. \quad (7)$$

Lance et al. (2016) evaluated the performance of their C9 and C10 SFIs via simulation using six population models that varied in complexity. For each of the six models, the population model was used to generate 1000 data sets where SM_{target} was the generating model containing paths that are either zero or non-zero in the population. This was done for each of four sample sizes: N = 100, 200, 500, 1000. Lance et al. (2016) introduced structural misspecifications by freeing 1 or 3 parameters (i.e., -1 or -3 df relative to $df_{SM_{target}}$) that are zero in the population, corresponding to C10 tests. In a similar vein, C9 tests were examined, where 1 or 3 parameters were fixed to zero that are non-zero in the population (i.e., +1 or +3 df relative to $df_{SM_{target}}$). Therefore, seven models were estimated to determine the performance of their proposed structural fit indices to detect structural misspecifications and global AFIs with a C10 interpretation (RMSEA and SRMR) and a C9 interpretation (CFI and TLI). Statistical power was assessed at different levels: 0.95, 0.975, 0.99, 1.0, with Type II error (β) rate equal to 1 minus power. For the global fit indices - SRMR, RMSEA, CFI, TLI, the cut-offs from Hu and Bentler (1999) were utilized.

Lance et al. (2016) found that both the $\frac{T}{df}$ (perDF) and $T - df$ (NCP) indices performed well from both the C9 and C10 interpretations. By assessing the power of the proposed fit indexes to detect mild misspecified structural models (e.g., +1 or -1 df), Lance et al. (2016) recommended C9 ≥ 0.99 and C10 ≤ 0.01 as cut-offs. Using these cut-offs, Lance et al. (2016) compared their performance to other approaches to assessing structural model fit: $\triangle\chi^2$, with critical values based on $\alpha = 0.01$ and RMSEA-P based on a cut-off of 0.08. Averaging over all six population models, they concluded that their proposed fit indexes outperformed both $\triangle\chi^2$ and RMSEA-P; further, Lance et al. (2016) noted that RMSEA-P and $\triangle\chi^2$ tended to select under-parameterized models, whereas the opposite behavior was found for their C9 and C10 indices. This work also complements previous

research, which indicates that global fit indices should not be used to assess structural model fit (McDonald & Ho, 2002; Williams & O'Boyle Jr, 2011).

The simulation conducted by Lance et al. (2016) was limited in scope and lacking in some details with respect to the reporting of results. First, the population models contained few indicators per factor (e.g., 1 to 4) and all loadings were 0.80 (standardized). Second, their selection of Type II error rate was commensurate with the recommended cut-offs proposed by Hu and Bentler (1999); however, the conditions that informed these cut-offs differed from the conditions investigated by Lance et al. (2016). Related to cut-offs, they did not systematically vary the standardized loadings when determining the cut-off values for the C9 and C10 structural fit indices; therefore, the same behavior shown by Heene et al. (2011) regarding the Hu and Bentler cut-offs applies to those proposed by Lance et al. (2016). McNeish and Hancock (2018) conducted a simulation in which the only manipulated factor was the size of the standardized loadings and showed that SM-LV is prone to the problems associated with measurement quality. Third, Lance et al. (2016) collapsed their results over all six models, making it impossible to determine whether or not the performance of their C9 and C10 indices vary as a function of model complexity; nor is it possible determine the performance of RMSEA-P in this respect. It is hypothesized that models with many structural parameters will aid RMSEA-P, compared with models with fewer structural parameters.

Two-Stage Measures - SM-MV. The SM-MV approach relies on the estimation of two models and therefore, is a two-stage estimation approach. First, the CFA model (e.g., SM_{sat}) is estimated and information is extracted. The extracted information from the initial model is re-purposed as observed data in a path analysis (i.e., latent variables are treated as observed). The authors claim that their approach provides the opportunity to assess structural model fit without interference from the measurement model.

Procedure. Upon estimating SM_{sat}, the estimated latent covariance matrix ($\hat{\mathbf{\Phi}}$) is extracted and is subsequently used as input ($\mathbf{\Omega}_{sat}$) for a path analysis where the causal

relations from the target model are subjected to these model-implied moments. Estimation completes when the distance between the observed and the model-implied covariance matrix ($\min[\boldsymbol{\Omega}_{sat} - \boldsymbol{\Omega}_{target}]$) has been minimized, corresponding to the *structural discrepancy fit function*, $\tilde{F}_{target}$. When only the covariance structure is included it is

$$\tilde{F}_{target}(\hat{\boldsymbol{\Omega}}_{sat}, \hat{\boldsymbol{\Omega}}_{target}) = \ln \left|\hat{\boldsymbol{\Omega}}_{target}\right| + tr(\hat{\boldsymbol{\Omega}}_{sat}\hat{\boldsymbol{\Omega}}_{target}^{-1}) - \ln \left|\hat{\boldsymbol{\Omega}}_{sat}\right| - K, \tag{8}$$

When the mean structure ($\hat{\kappa}$) is included, $\tilde{F}_{target}$ is

$$\begin{aligned}\tilde{F}_{target}(\hat{\boldsymbol{\Omega}}_{sat}, \hat{\boldsymbol{\Omega}}_{target}; \hat{\kappa}_{sat}, \hat{\kappa}_{target}) = (\hat{\kappa}_{sat} - \hat{\kappa}_{target})\hat{\boldsymbol{\Omega}}_{target}^{-1}(\hat{\kappa}_{sat} - \hat{\kappa}_{target})^t + \\ \ln \left|\hat{\boldsymbol{\Omega}}_{target}\right| + tr(\hat{\boldsymbol{\Omega}}_{sat}\hat{\boldsymbol{\Omega}}_{target}^{-1}) - \ln \left|\hat{\boldsymbol{\Omega}}_{sat}\right| - K,\end{aligned} \tag{9}$$

In Equations 8 and 9, K corresponds to the number of LVs estimated in SM_{sat}. To estimate SFIs from the incremental perspective, it is necessary to estimate the structural discrepancy fit value for the baseline (or null model). When only the variance structure is included, $\tilde{F}_{null}$ is

$$\tilde{F}_{null}(\hat{\boldsymbol{\Omega}}_{sat}, \hat{\boldsymbol{\Omega}}_{null}) = \ln \left|\hat{\boldsymbol{\Omega}}_{null}\right| + tr(\hat{\boldsymbol{\Omega}}_{sat}\hat{\boldsymbol{\Omega}}_{null}^{-1}) - \ln \left|\hat{\boldsymbol{\Omega}}_{sat}\right| - K. \tag{10}$$

When the mean and covariance structures are included, $\tilde{F}_{null}$ is

$$\begin{aligned}\tilde{F}_{null}(\hat{\boldsymbol{\Omega}}_{sat}, \hat{\boldsymbol{\Omega}}_{null}; \hat{\kappa}_{sat}, \hat{\kappa}_{null}) = (\hat{\kappa}_{sat} - \hat{\kappa}_{null})\hat{\boldsymbol{\Omega}}_{null}^{-1}(\hat{\kappa}_{sat} - \hat{\kappa}_{null})^t + \\ \ln \left|\hat{\boldsymbol{\Omega}}_{null}\right| + tr(\hat{\boldsymbol{\Omega}}_{sat}\hat{\boldsymbol{\Omega}}_{null}^{-1}) - \ln \left|\hat{\boldsymbol{\Omega}}_{sat}\right| - K,\end{aligned} \tag{11}$$

Estimate pseudo fit statistics and indices. In the same manner as before, a *pseudo test statistic*, $\tilde{T}$, can be estimated for both the target and the null structural model using $\tilde{F} * N$, denoted as $\tilde{T}_{target}$ and $\tilde{T}_{null}$, respectively.

$$\nu_{target} = \frac{K(K+1)}{2} - t_{target}; \tag{12}$$

To determine the number of degrees of freedom for the target model (ν_{target}) we utilize Equation 12 above, where t_{target} corresponds to the total number of free parameters in

the path analysis. Hence, ν_{null} is determined as:

$$\nu_{null} = \frac{K(K+1)}{2} - K, \tag{13}$$

this is due to the off diagonals of $\tilde{\mathbf{\Omega}}_{null}$ being set to zero.

Using these pieces of information, we can construct structural fit indices (e.g., $RMSEA_S$) as before and utilize them for descriptive purposes (Hancock & Mueller, 2011). The estimate of $\tilde{T}_S$ is sample based and its global equivalent is T_{ML} with ν degrees of freedom. The sample estimate of the non-centrality parameter is $\lambda_S = \tilde{T} - \nu_{target}$. d_S is then estimated by $\frac{\tilde{T}-\nu_{target}}{N-1}$.

The sole empirical investigation of the SM-MV approach was conducted by McNeish and Hancock (2018), who illustrated that the Lance et al. (2016) SM-LV approach and their prescribed cut-offs are prone to the reliability paradox. Further, McNeish and Hancock (2018) showed that the SM-MV approach can aid the performance of the SM-LV SFIs in the context of varying levels of measurement quality. Based on their reporting, it appears that the SM-MV SFIs might also be affected by measurement quality; specifically, McNeish and Hancock (2018) reported median and standard deviation estimates for SM-MV SFIs that indicate a clear pattern: as measurement quality decreases, standard deviation estimates increase. Without knowing the mean estimates for the SFIs, it is impossible to fully understand the implications of their standard deviations. For instance, the median values for SRMR and RMSEA are approximately 0.053 and 0.28, respectively; therefore, half of the estimates must fall below these values. Because the SRMR and RMSEA are bounded by zero, their sampling distributions must be positively skewed, which affects their statistical power to detect a structural misspecification. This same pattern was observed by McNeish and Hancock (2018) when using the SM-MV approach to aid in the estimation of the SM-LV SFIs. The selective reporting by McNeish and Hancock (2018) masks the effect of two-stage estimation and its consequences are not immediately known. It is hypothesized that as sample size decreases, the standard deviation estimates for the SM-MV SFIs will increase. Interestingly, neither McNeish and Hancock (2018) or Hancock and Mueller (2011) warn researchers about the uncertainty that surrounds $\hat{\Phi}$ or $\hat{\alpha}$. Due to the

two-stage nature of the SM-MV approach, unintended consequences were hypothesized to surface. Specifically, given poor measurement quality and the correct model in the population, the SM-MV SFIs will not reliably indicate the model fits the data.

Misfit in Mean Structure

As previously noted, F_{ML} serves as the basis from which nearly all of the AFIs are calculated. As Equation 6 illustrates, the estimation of F_{ML} takes the mean structure into account, therefore, T and all AFIs that depend on T should be capable of detecting misfit in the mean structure (Wu & West, 2010). The majority of the research that has been done on the performance of measures of fit with misspecified mean structures has been in two areas: latent growth curve models (LGC) and multiple group analysis (MG-CFA/SEM). Both of these modeling contexts provide unique opportunities for evaluating the mean structure. First, the mean structure is commonly over-identified (i.e., through constraints placed on mean parameters). Second, the specification of the mean structure in the baseline model (i.e., used to estimate incremental AFIs) is in accordance with the null hypothesis at hand. Below these modeling contexts are detailed.

Latent Growth Curve Models. The goal of modeling a latent growth curve is to estimate change in a construct over time. This is done by fixing elements in Λ to represent the hypothesized growth (i.e., no growth or linear growth). Similarly, the elements in τ are fixed to zero thereby transmitting all of the mean information to the latent space. Only elements in Θ are freely estimated. Therefore, the elements in α are the marginal means or the population estimates of the growth parameters that define the functional form (i.e., the expected value at the intercept and the expected rate of change at the intercept). The diagonal elements in Φ represent the *between-person* variance around their respective growth parameters (i.e., variability at the intercept and rate of change); the off-diagonal estimates describe the degree to which the growth parameters are related. Finally, *within-person* variability is represented by Θ and corresponds to variance unaccounted for by the growth process (i.e., measurement

error).

Mean structure degrees of freedom. Because the elements of Λ and τ are not freely estimated, the mean and covariance structures are over-identified. For instance, consider a longitudinal study in which measurements on a scale were taken at 5 equally spaced occasions across time (i.e., p = 5). To test the null hypothesis of no change over time, an intercept only model is estimated. The parameters that are freely estimated are: the latent mean (α_{11}) and variance (Φ_{11}), while constraining the diagonal elements of Θ to be the same. As a result, a single mean parameter is estimated and two variance parameters are estimated. In terms of the amount of information available, 5 unique pieces exist for the mean structure (4 degrees of freedom) and 15 unique pieces of information are available for the covariance structure (13 degrees of freedom) - totaling 17 degrees of freedom. To test alternative trajectories, additional growth parameters can be modeled. In the event a linear trend is modeled, an additional mean parameter is estimated, whereas, an additional 2 covariance parameters must be estimated (i.e., the covariance between the intercept and slope, Φ_{01}). As a result of fitting the linear trend, the mean structure contains 3 degrees of freedom, compared to 11 degrees of freedom in the variance structure. This illustrates the unique role of the mean structure and how misfit can more readily occur in this structure.

Types of misfit. In sum, misfit can occur in both the mean and variance structures. For instance, the marginal means (α) could be misspecified by freely estimating a manifest intercept (τ), rather than fixing it to zero, which would alter the growth trajectory. The between-person variance-covariance matrix (Φ) could be misspecified by fixing the covariance between the intercept and rate of change growth parameters to zero, when it is non-zero in the population; with respect to the within-person variance-covariance matrix, homogeneous errors could be estimated when the measurement error is different across time in the population (Wu & West, 2010).

To test whether the functional form is correct, growth terms are added to the model, as separate LVs with their Λ values fixed according to the form (e.g., quadratic: 0, 1, 4, 9). Afterwards, difference estimates $\triangle T_{\triangle df}$ are calculated and a $\triangle\chi^2$ test is performed.

The null hypothesis states that the more parsimonious model provides adequate fit; therefore, if the null hypothesis is rejected, the more parameterized model is favored. In a similar fashion, the decision on whether to estimate a variance for a given growth parameter can be made. Note that these tests are all related to structural fit, as the only measurement parameter being estimated is Θ. For the evaluation of model fit via incremental fit indices to be valid, the correct baseline model must be specified.

Baseline model. Earlier, incremental indices were introduced and the independent null model was presented; however, in the context of LGC models, the independent null model should not be used to calculate incremental fit indices in this modeling context (Wu & West, 2010; Wu, West, & Taylor, 2009). Of great importance, the baseline model should represent the null hypothesis and in the context of LGC models this corresponds to an intercept only model, which was presented above. This is because when we estimate a LGC model, we hypothesize that there is growth on the measure over time. The intercept only (or baseline) model assumes homogeneous errors and the fit information from this model is retained, affording the ability to correctly estimate incremental AFIs, making it possible to make an appropriate judgment on model fit. In the case of LGC models, the specification of the mean structure in the baseline model is straightforward, however, this is not the case for all models.

Performance of AFIs and exact fit. Wu and West (2010) conducted a comprehensive simulation to evaluate the performance of T, RMSEA, SRMR, CFI, and TLI in the context of misspecified LGC models. Specifically, they tested four types of misspecifications. Misspecification of the marginal mean structure was accomplished by omitting the quadratic term from the model. Misspecification of the model-implied variance-covariance structure was accomplished by either (a) setting the variance for the quadratic term to zero, (b) setting the covariance between the intercept and linear term to 0, (c) constraining the residual variances to be equal, or (d) by fixing the autoregressive term to zero when it was non-zero in the population. Systematically, they crossed these misspecification types and examined the impact of sample size (N = 125, 250, 500, 1000) and degree of misspecification (slight, moderate, and severe) on the

performance of the AFIs in detecting the misspecifications.

First, they estimated the population models and found that all indices agreed that the model fit the data well over the 1100 replications across all sample sizes: T(df = 9); TLI and CFI were near 1.0; and SRMR and RMSEA were near 0.0. When only the covariance structure was misspecified, the degree of misspecification affected T, RMSEA, CFI, and TLI more than SRMR. The effect of sample size was found to impact the performance of T (increased with sample size) and SRMR (decreased with sample size), whereas CFI, TLI, and RMSEA were not affected. With respect to the type of misspecification, SRMR was highly impacted, while T, CFI, TLI, and RMSEA were impacted less. Wu and West (2010) found that across the AFIs and T, all were least sensitive to the misspecification resulting from fixing the variance for the quadratic term to zero, whereas all were most sensitive to fixing the covariance between intercept and rate of change to zero.

When only the marginal mean structure was misspecified, the degree of misspecification impacted all of the measures, with SRMR and T being affected less than the others. T and SRMR were again impacted by sample size. The CFI, TLI, and RMSEA were considerably less sensitive to misspecification in the mean structure than misspecification of the covariance matrix. When both the marginal mean and covariance structures were misspecified, it was found that only RMSEA and CFI were affected; the rest of the measures were impacted more by one type of structure misspecification over the other. When the quadratic term and its variance were omitted (estimating a linear LGC model), all measures were most sensitive to the mean structure. In sum, Wu and West (2010) recommend the use of RMSEA, CFI, and TLI over T and SRMR. The recommendation of CFI and TLI is conditional on whether the correct baseline model is used.

Multiple Group Models. When the goal is to investigate whether groups differ with respect to the latent mean(s) (α) or the latent variances-covariances (Φ), it must be shown that the measure possesses the quality of measurement invariance across the groups of interest. To this end, MG-CFA/SEM models are employed and a series of four

models are estimated:

1. The *form invariant* model tests whether or not the factor pattern (e.g., fixed and free parameters) is the same for the two groups while identifying each group in the same fashion (e.g., marker variable). This model produces unique estimates for all free parameters for each group. Given the same number of MVs (P = 25) and LVs (K = 5) as described above, the number of free parameters for the form invariant model is 170 (85 per group) with 530 degrees of freedom (265 per group), due to there being mean and variance information available for each group.

2. The *metric invariant* model tests whether the Λ parameters (i.e., unstandardized regression estimates) can be constrained to be the same across groups; as a result, 20 degrees of freedom are gained (i.e., 20 rather than 40 free Λ estimates).

3. The *scalar invariant* model tests whether the τ estimates can be constrained to be the same across groups, which represents the expected value of the MV when the LV is zero; as a result, 20 degrees of freedom are gained in the mean structure (i.e., 20 rather than 40 free τ elements).

4. The *strict invariant* model tests whether or not the proportion of variance unexplained by the LVs is the same for the two groups.

Typically, strict invariance is uncommonly sought in practice and the majority of the research on establishing measurement invariance is focused on the form, metric, and scalar models. Without meeting form and metric invariance, comparisons on latent variances-covariances are not meaningful; further, if scalar invariance is not met, comparisons on the latent means are also not meaningful.

Impact on degrees of freedom and types of misfit. As we move from the form invariant model to the metric invariant model, we gain degrees of freedom in the covariance structure; and when moving from the metric invariant model to the scalar invariant model, we gain degrees of freedom in the mean structure. However, as we gain degrees of freedom, this introduces the opportunity for mean and/or covariance misfit to occur.

Recall that the form invariant model results in unique estimates for each group. As a result, the form invariant model usually is the best fitting model compared with the metric, scalar, and strict invariant models. It follows then, that as sets of parameters (e.g., Λ for metric invariance) are constrained to be the same across groups, model misfit may occur - representing measurement non-invariance (i.e., the measure being examined functions differently across groups). Therefore, when assessing metric invariance, the covariance structure can be misspecified as a result of constraining Λ to be the same across groups, or the mean structure can be misspecified as a result of constraining τ to be the same across groups when one or more elements in Λ and τ, respectively, are different across the two populations. Because T is affected by sample size, it has been suggested to use the change in AFIs ($\triangle$ AFIs) to assess whether a significant amount of misfit was produced from constraining a set of parameters (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008); if the decrease in fit is negligible, then the constraint is said to be tenable. The seminal work of Chen (2007) is detailed below; however, a discussion on baseline models is warranted first.

Baseline models. As noted earlier, the selection of the baseline model has ramifications for the utility and performance of incremental fit indices, such as CFI and TLI. In the context of multiple group CFA/SEM models, we are interested in group differences, and the appropriate baseline model must be estimated to reflect the null hypothesis (i.e., groups do not differ from one another). To this end, several null models have been proposed.

First, Little et al. (2007) described a null model where MVs are entered into the model as observed latent variables. In order to model an MV as an observed latent variable, it is necessary to fix factor loadings to 1.0 ($\Lambda = 1$) and manifest residuals to 0.0 ($\Theta = 0$), which forces all of the variance information observed into the latent space. In a similar vein, the manifest intercepts are fixed to 0.0 ($\tau = 0$), which forces the observed mean information into the latent space. As a result, the diagonal elements of Φ are estimated and constrained to be the same across groups, and the off-diagonal elements are fixed to 0.0. With respect to the means, all elements of α are freely estimated and constrained

to be the same across groups. This then models the null hypothesis that groups do not differ with respect to either the mean or variance structure.

More recently, Lai and Yoon (2015) proposed an alternative CFI (referred to as $CFI_{=}$) that builds on the approach proposed by Rigdon (1998). $CFI_{=}$ carries the assumption that the factor structure is known prior to investigating measurement invariance, as evidenced by the specification of the form invariant model (i.e., pattern of fixed and free parameters). Therefore, the baseline model estimates a common correlation for MVs that load onto the same factor and a common factor correlation if there is more than one LV measured. With respect to the mean structure, τ is freely estimated and constrained to be the same across groups, and α is fixed to 0.0.

$\triangle AFIs$. Cheung and Rensvold (2002) conducted a large simulation study to assess Type I error rates of $\triangle AFIs$ when testing for both measurement and structural invariance. From this study, Cheung and Rensvold (2002) derived a cut-off of 0.01 for $\triangle CFI$ and 0.02 for McDonald's measure of centrality $\triangle Mc$. In the context of assessing measurement noninvariance (metric, scalar, and strict) with an α of 0.01, Chen (2007) derived cut-offs of 0.005 and 0.01 for $\triangle CFI$ and $\triangle Mc$, respectively. Meade et al. (2008) assessed AFIs in the context of both measurement invariance and noninvariance, ultimately they found that the cut-offs proposed by Cheung and Rensvold (2002) for $\triangle CFI$ and $\triangle Mc$ were adequately powered to detect measurement noninvariance. Further, Meade et al. (2008) investigated the effect construct reliability had on $\triangle CFI$ and $\triangle Mc$. They found that power levels for $\triangle CFI$ were unaffected and suggested a cut-off of 0.002, whereas power levels varied for $\triangle Mc$, and they could not recommend a standard cut-off. When the amount of misfit is significant (surpasses $\triangle AFI$), this is an indication that at least one factor loading (metric invariant model) or at least one intercept (scalar invariant) must be freely estimated across groups. After removing the constraint on one or more of the offending elements in Λ or τ and the amount of misfit is negligible, partial invariance is established.

Given that partial metric invariance is established, groups can be compared with respect to latent variances-covariances (Byrne, Shavelson, & Muthén, 1989; Shi, Song,

& Lewis, 2017), and if the partial invariance model is correct, these comparisons are unaffected (French & Finch, 2016). Typically, an omnibus test of the latent variances is conducted where constraints are placed on all elements in Φ; if misfit results, just the diagonal elements of Φ are tested. When this constraint is found to be tenable, the off-diagonal elements can be tested; however, if the latent variances cannot be constrained across groups, a standardization procedure (via phantom variables) can be utilized to test for differences in the correlations between groups. Next, the latent mean structure is typically investigated as long as partial scalar invariance holds across groups (Byrne et al., 1989; Shi et al., 2017). Testing occurs in a similar fashion, where an omnibus test is performed, constraining all of the elements in α. Misfit while testing for structural invariance indicates a difference in the population between the groups being examined. Typically, $\triangle\chi^2$ is used to test whether or not a given constraint on structural parameters is tenable (Anderson & Gerbing, 1988; Thompson & Green, 2006). This is due to AFIs being less sensitive to structural misspecifications, as demonstrated by Fan and Sivo (2009).

Fan and Sivo (2009) investigated the performance of $\triangle AFIs$ in their ability to detect latent mean differences. Fan and Sivo (2009) varied the number of factors (2, 3, or 4), the number of indicators per factor (p:f; 2, 4, or 6), and the ratio of sample size and number of indicators, (n:p; 20/1, 40/1, or 60/1), and 7 levels of latent mean difference (d = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, or 0.8). With respect to measurement quality, all factor loadings were 0.7 in the population. First, they derived threshold values for nine AFIs and settled on 0.02 for the $\triangle CFI$, $\triangle RMSEA$, and $\triangle Mc$ using the population model where d = 0.0. They proceeded to fit an unconstrained model where only the latent means were freely estimated across the two groups, followed by a constrained model to investigate the performance of their proposed thresholds. Fan and Sivo (2009) found that $\triangle Mc$ outperformed the rest of the AFIs; as models became larger, statistical power for all of the $\triangle AFIs$ diminished to zero, with $\triangle Mc$ being the exception as its statistical power remained constant or increased. When the latent mean difference was 0.5, the statistical power for $\triangle CFI$ and $\triangle RMSEA$ was always greater for a two-factor

model compared with a four-factor model. Power levels were high when p:f was 2 (power = 100); however, when p:f reached 6, power levels dropped to zero or near zero. On the other hand, $\triangle Mc$ power levels were consistently at 100. Ultimately, Fan and Sivo (2009) concluded that $\triangle AFIs$ should not be used to make inferences about difference in latent means between two populations; instead, $\triangle\chi^2$ should be used.

Comparison of baseline models - $\triangle CFI$. Lai and Yoon (2015) conducted several simulations that varied sample size (N = 200 or 500 per group); number of LVs (1 or 2 - with $\Phi_{2,1} = 0.3$), each with six indicators; and the number of non-invariant indicators (1 or 2). In the simulation to assess metric invariance, Λ_{G1} was set to 0.7 and Λ_{G2} varied based on the degree of non-invariance: $\triangle\Lambda = 0.0, 0.2, 0.4$. In the simulation to assess scalar invariance, $\tau_{G1} = 0.0$ and the degree of non-invariance was varied: $\triangle\tau = 0.0, 0.2, 0.5$.

When investigating metric non-invariance, Lai and Yoon (2015) found that $CFI_{=}$ outperformed both $\triangle CFI$ and CFI with the independent null model as its baseline model. When two factor loadings were non-invariant ($\triangle\Lambda = 0.4$), CFI had a mean estimate of 0.957-0.958 over the simulation conditions and indicated poor fit in only 14.8 to 32.8 percent of the replications. With respect to $\triangle CFI$, when the cut-off proposed by Cheung and Rensvold (2002) was used, $\triangle CFI$ performed better across all simulation conditions; when the cut-off proposed by Meade et al. (2008) was used, slightly higher power to detect the misfit was obtained, but the Type I error rate was higher compared with $CFI_{=}$ (0.16 and 0.10, respectively). Lai and Yoon (2015) also investigated AFIs from the absolute perspective and found that RMSEA was affected by model size and performed worse as model size increased. Non-invariance was detected in only 21 percent of the replications across simulation conditions. The residual based SRMR was found to perform worse than RMSEA and only detected non-invariance 6 percent of the time. When investigating misfit based on $\triangle\chi^2$, Lai and Yoon (2015) found that this approach was able to detect the non-invariance in 86 percent of the replications.

With respect to scalar invariance, CFI with the independent null model as its baseline had mean estimates between 0.992 and 0.942 for the non-invariant simulation conditions

and at best detected misfit in 83.8 percent of the replications when sample size was 500 per group with 2 LVs and two non-invariant intercepts ($\triangle\tau = 0.5$). $CFI_{=}$ reached appropriate power levels when $\triangle\tau$ was 0.2 for two MV intercepts and detected non-invariance in 68.2 to 93.4 percent of the replications, per condition; using the Meade et al. (2008) cut-offs, non-invariance was detected in 71.2 to 95.8 percent of the replications per condition. When there was only one non-invariant MV intercept, the $\triangle\chi^2$ based on the Meade et al. (2008) cut-offs outperformed $CFI_{=}$ and detected the non-invariance 55.8 to 81.8 percent of the replications compared with 46.4 to 55.2 percent of the replications. With respect to the AFIs from the absolute perspective, the RMSEA and SRMR performed poorly across all conditions compared with $\triangle CFI$ and $CFI_{=}$. On the other hand, the $\triangle\chi^2$ performed similarly to $\triangle CFI$ and $CFI_{=}$ with better Type I error rates.

The pattern that emerges shows that AFIs and augmented AFIs across the board had a harder time detecting non-invariance (or misfit) in the mean structure. As Lai and Yoon (2015) illustrated, the CFI estimated based on the independent null model is clearly unacceptable for detecting measurement non-invariance and showed that detection of non-invariance can be improved by using $\triangle CFI$ and the cut-offs proposed by Meade et al. (2008). The modified baseline CFI ($CFI_{=}$) yields better Type I error rates than $\triangle CFI = 0.002$ and appears to perform as well as $\triangle\chi^2$ with respect to power for detecting non-invariant factor loadings and manifest intercepts.

Measurement quality. Kang, McNeish, and Hancock (2016) conducted two simulation studies that investigated the impact of measurement quality in the context of multiple group models. The first simulation focused on measurement invariance. The purpose of the measurement invariance study was to empirically determine cut-off values for $\triangle Mc$ and $\triangle CFI$, moving from the form invariant to the metric invariant (data generation) model, adhering to an α of 0.01 and 0.05. Cheung and Rensvold (2002) recommend $\triangle CFI = -0.01$ and $\triangle Mc = -0.02$ to indicate invariance. In this study, the standardized loadings varied from 0.4 to 0.95, by steps of 0.05; group sample size varied (n = 100, 200, 300, 600, 1000); and the number of indicators varied between 3 or 5;

using a single LV with two groups. Kang et al. (2016) determined that under these simulation conditions $\triangle Mc$ outperformed $\triangle CFI$ because it was not affected by the magnitude of the loadings, the number of indicators per factor, or sample size. Therefore, Kang et al. (2016) recommended $\triangle Mc_{\alpha=0.01} = -0.007; \triangle Mc_{\alpha=0.05} = -0.01$. The second simulation focused on structural invariance. This second study used a population model containing three exogenous and endogenous LVs, each measured by three indicators. Kang et al. (2016) generated group differences in the three structural paths that corresponded to standardized differences (d) of 0.7, 0.4, and 0.2, between the two groups (n = 1000). Similar to the measurement invariance study, the factor loadings were the only aspect that varied, aside from the standardized differences. $\triangle$ fit indices were calculated moving from the unconstrained (population) to constrained (misspecified) model and were ultimately compared with the cut-offs suggested by Cheung and Rensvold (2002). Kang et al. (2016) determined that $\triangle Mc$ and $\triangle CFI$ are both affected by the magnitude of the loadings. When detecting a large misspecification (d = 0.7), $\triangle CFI$ required loadings to be 0.55 or larger, whereas, $\triangle Mc$ always detected the misspecification. In the context of a moderate misspecification (d = 0.4), $\triangle Mc$ performed well when $\Lambda = 0.55$, whereas, $\triangle CFI$ only detected the misfit when $\lambda = 0.85$. In the context of a small misspecification, only $\triangle Mc$ could detect the misfit and required $\Lambda = 0.75$. Therefore, Kang et al. (2016) determined that $\triangle Mc$ outperformed $\triangle CFI$. Aside from the effect measurement quality had on $\triangle CFI$ and $\triangle Mc$, they found the standard errors of the structural parameters were also affected, which in turn affected the test statistic. Specifically, Kang et al. (2016) showed that when the structural parameter is 0.7 in the population and measurement quality is high ($\Lambda = 0.95$), the standard error and z score are 0.018 and 38.46, respectively, for the parameter, whereas, when measurement quality is low ($\Lambda = 0.40$), its standard error and z-score are 0.096 and 7.33, respectively. This finding is similar to that of Hancock and Mueller (2011).

As research illustrates, researchers need the ability to parse global fit to evaluate structural model fit in isolation. Evaluation of structural model misfit based on AFIs is

hampered due to the size of the measurement model and the effect of measurement quality on the standards we use to decide whether to retain or dismiss a model. The SM-MV and SM-LV approaches are separate approaches to this problem; however, these approaches are both in their early stages of investigation and validation. The SM-LV is quite recent and has only been empirically examined once; the SM-MV approach was proposed seven years ago, but less empirical work has been conducted on it than the SM-LV approach. The only instance of an empirical investigation of the SM-MV approach was in response to the introduction of the SM-LV, as previously mentioned. Further, there is a lack of evidence regarding the effect two-stages estimation has on the SM-MV approach. Related, it appears an opportunity to evaluate the two-stage estimation approach of Hancock and Mueller (2011) in the context of multiple group models was missed by Kang et al. (2016).

Clearly, more empirical work is needed to determine the merits of these approaches for evaluating structural model misfit. The purpose of this dissertation is to systematically examine the SM-MV and SM-LV approaches over varying simulation conditions in the context of group differences. I will add to the literature by investigating the merits of SM-MV and SM-LV in four novel ways. I will examine the relative performance of the approaches in detecting: (1) misfit in the mean structure; (2) misfit in the covariance structure; (3) misfit in both the mean and covariance structure; and (4) their performance in multiple group models. Additionally, I will compare the SM-MV and SM-LV approaches to RMSEA-P, as their relative merits are unknown.

In the social sciences and education, it is common for researchers to model group differences using LV models, and the parameters that are germane to such research questions are structural. It is imperative that researchers understand what constitutes a reasonable approximation based on values of RMSEA-P and the structural fit indices that result from the SM-MV and SM-LV approaches. In this vein, it is also important for researchers to understand the statistical power of these structural fit indices to detect true population differences between groups that correspond to a small, medium, or large effect size. Further, knowing the performance of these structural measures of fit

over varying model and sample characteristics to detect different types of misspecifications will inherently lead to stronger inferences. This study provides such guidance for those who rely on multiple group models to answer their research questions.

Research Questions

The research questions addressed in this study were:

1. When the structural model is correctly specified, what is the sampling distribution of structural measures of fit (RMSEA-P, SM-MV, and SM-LV) and how do varying levels of p:f, measurement quality, sample size and characteristics (balanced or unbalanced) affect the distribution of the selected measures of structural fit? Relatedly, to what extent is the SM-MV approach affected by two stages of estimation?

2. What is the relative performance of the structural fit measures ($\triangle\chi^2$, RMSEA-P, SM-MV, and SM-LV) to detect population differences in either the mean or covariance structure that correspond to a small, medium, or large effect size?

   Do varying levels of p:f, measurement quality, sample size and characteristics (balanced or unbalanced) moderate the relative performance of these structural measures of fit?

3. What is the relative performance of the structural fit measures ($\triangle\chi^2$, RMSEA-P, SM-MV, and SM-LV) to detect population differences that simultaneously exist in the mean (small, medium, or large) and covariance (small, medium, or large) structures?

   Do varying levels of p:f, measurement quality, sample size and characteristics (balanced or unbalanced) moderate the relative performance of these structural measures of fit?

4. How do structural fit measures ($\triangle\chi^2$, RMSEA-P, SM-MV, and SM-LV) compare to global fit indices in their ability to detect a structural model that is incorrect in the population?

It was hypothesized that the SFIs estimated using the SM-MV approach will be negatively impacted by measurement quality, especially when the amount of information available per MV decreases. It was expected that the SFIs that result from the SM-LV approach also will be affected by measurement quality; however, it was hypothesized that it will outperform SM-MV when the subsample size decreases. With respect to the RMSEA-P, it was hypothesized that it will be negatively impacted when model size is small and will perform better as model size increases. It also was hypothesized that RMSEA-P will perform similarly to $\triangle\chi^2$, but will be less affected by sample size than χ^2.

3. Method

This study sought to achieve two goals in the context of multiple group SEM. The first goal was to understand the sampling distribution of proposed structural measures of fit (Type I Error simulation). The second goal was to determine the relative power of the proposed structural measures of fit (Power simulation). In the Type I Error simulation cut-off values were empirically derived and were utilized in the subsequent Power simulation to evaluate the merit of the proposed approaches in their ability to detect structural model misfit. In this chapter, I detail methods that remained uniform across the two simulations.

Data Generation

The model utilized in this study was motivated by R. MacCallum (1986) and is shown as a single group model in Figure 1. Lance et al. (2016) used a similar model to establish cut-off values for their C9 and C10 indices and it later served as the model from which McNeish and Hancock (2018) compared the SM-MV and SM-LV approaches. Earlier, this model was used to investigate the performance of parsimony indices (Williams & Holahan, 1994). As Figure 1 illustrates, the model contains 5 LVs - 3 of which are exogenous (X1, X2, and X3) and 2 of which are endogenous (Y1 and Y2), and possesses both direct and indirect effects among the LVs. In this mediation model, Y1 fully mediates the effect of both X1 and X3 on Y2 and partially mediates the effect of X2 on Y2. In all instances, manifest and latent variables were generated from the multivariate normal distribution in the R environment (R Core Team, 2017) using the simsem package (Pornprasertmanit, Miller, & Schoemann, 2016). Data was generated according to full measurement and structural invariance (i.e., groups did not differ from one another in the population).

Structural Model. The total variance for all of the latent variables was fixed to one, which allowed the elements of $\mathbf{B}$ to be interpreted as standardized regression parameters and the off-diagonal elements of $\mathbf{\Psi}$ to represent latent correlations. To accomplish this, it was necessary to solve for the latent variance unexplained by the

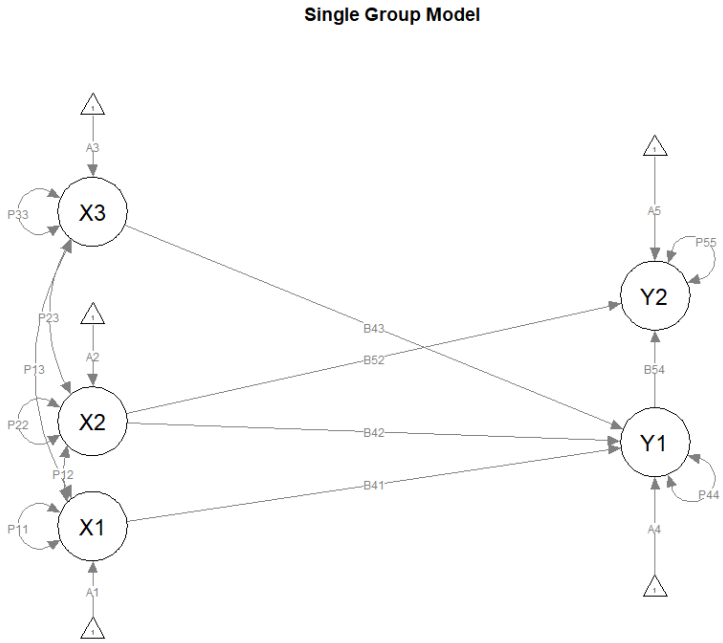

Figure 1. Path Diagram: Data Generating Model

Single group model with no misspecifications. Exogenous latent variables: X1, X2, X3. Endogenous latent variables: Y1 and Y2.

system of equations for the endogenous latent variables Y1 ($\mathbf{\Psi}_{[44]}$) and Y2 ($\mathbf{\Psi}_{[55]}$). The variance explained by the system of equations for Y1 was solved for and its result was subtracted from 1.0 to determine the population value for $\mathbf{\Psi}_{[44]}$:

$$\Phi = B_{41}^2 + B_{42}^2 + B_{43}^2 + 2(B_{41} * \Psi_{12} * B_{42}) + 2(B_{41} * \Psi_{13} * B_{43}) + 2(B_{42} * \Psi_{23} * B_{43})$$

$$\mathbf{\Psi}_{[44]} = 1 - \Phi$$

$$\mathbf{\Psi}_{[44]} = 0.622 \tag{14}$$

Afterwards, $\mathbf{\Psi}_{[55]}$ was set to 0 and the variance explained by the system of equations was accomplished via:

$$\mathbf{\Phi} = (\mathbf{I} - \mathbf{B})\mathbf{\Psi}(\mathbf{I} - \mathbf{B})^t \tag{15}$$

As before, the variance explained by the system of equations for Y2 (Φ_{55}) was

subtracted from 1.0 and this value was the population value for ψ_{55}. The structural matrices $\mathbf{B}$ and $\mathbf{\Psi}$ used for data generation are presented in Tables 1 and 2, respectively; recall that $\mathbf{I}$ is an identity matrix of the same order as $\mathbf{B}$ and $\mathbf{\Psi}$ and therefore is not shown. The latent intercepts (ν) are a function of $\mathbf{B}$ and α (See Equation 16).

Table 1

***B**, Matrix of Latent Regressions*

	X1	X2	X3	Y1	Y2
X1	0	0	0	0	0
X2	0	0	0	0	0
X3	0	0	0	0	0
Y1	0.3	0.3	0.3	0	0
Y2	0	0.4	0	0.3	0

Table 2

$\mathbf{\Psi}$, *Matrix of Latent Variances and Disturbances*

	X1	X2	X3	Y1	Y2
X1	1	0.200	0.200	0	0
X2	0.200	1	0.200	0	0
X3	0.200	0.200	1	0	0
Y1	0	0	0	0.622	0
Y2	0	0	0	0	0.649

$$\nu = (\mathbf{I} - \mathbf{B})\alpha \tag{16}$$

The population values for the latent means were: $\alpha = [0, 0, 0, 1.5, 1.2]$ and with $\mathbf{B}$ resulted in the latent means and intercepts shown in Table 3.

Measurement Model. The population values for the measurement model matrices (e.g., $\mathbf{\Lambda}$) varied across simulation conditions to systematically investigate the effect of measurement quality on the evaluation of structural model fit. Across conditions, the

Table 3

ν, Vector of Latent Means and Intercepts

X1	X2	X3	Y1	Y2
0	0	0	1.5	0.75

total variance for all manifest variables was fixed to unity. Therefore, the population values for the manifest residuals (i.e., the diagonal matrix $\mathbf{\Theta}$) were determined as a function of Λ. Therefore, when measurement quality was poor (i.e., elements of $\mathbf{\Lambda}$ were 0.4), the diagonal elements of $\mathbf{\Theta}$ were set to $1 - 0.4^2 = 0.84$. The manifest intercepts (τ) were set to 0 in the population.

Procedure

After data were generated for a given replication, a total of six models were estimated. Upon fitting these six models, it was possible to calculate the measures of fit of interest (see Measures section below). The first model to be fitted was a baseline model in order to correctly estimate incremental fit indices in the context of multiple group models. Specifically, a variance and a mean are estimated for each observed variable and constrained to be the same across groups. For all latent variable models, the measurement model reflected both metric and scalar invariance. Therefore, the freely estimated elements in $\mathbf{\Lambda}$ and τ were constrained to be the same across groups. The latent variable models estimated were the structural null (SM_{null}), structural target (SM_{target}), and saturated structural (SM_{sat}) models. These models are detailed below.

- SM_{null}: This model estimates latent covariances among the exogenous latent variables (e.g., X1, X2, X3); however, these estimates are not constrained across groups. The relationships between the exogenous latent variables and the hypothesized endogenous latent variables (Y1 and Y2) were set to zero for both groups. In terms of the latent variances and means, these parameters were freely estimated across groups. With respect to the measurement model, constraints across groups were placed to model metric and scalar invariance, respectively.

- SM_{target}: This model was the data generating model, therefore, all of the correct structural paths were estimated and constrained to be the same across groups (i.e., paths $X1 \rightarrow Y2$ and $X3 \rightarrow Y2$ were fixed to zero). As a result, both measurement and structural invariance was modeled.

- SM_{sat}: The saturated structural model is akin to the correlated factors model. Therefore, all latent parameters (e.g., α and $\mathbf{\Phi}$) are estimated with no constraints placed on them across groups. With respect to the measurement model, both metric and scalar invariance is modeled.

Using information (e.g., latent means and variances-covariances) from SM_{sat}, path analysis was employed. Specifically, the two path models that were estimated were the null ($PATH_{null}$) and the target ($PATH_{target}$); these are detailed below.

- $PATH_{null}$: In this model, the same pattern of free and fixed paths specified among the latent variables in SM_{null} was specified.

- $PATH_{target}$: This model utilizes the same pattern of free and fixed paths among the latent variables in SM_{target}.

All latent variable and path models were executed in R (R Core Team, 2017) using the lavaan (Rosseel, 2012) package. All models were fitted using the sem function utilizing its *mimic = "mplus"* option and were estimated with maximum likelihood. Latent variable models were identified using the marker variable approach and the first indicator per factor was chosen as the reference indicator.

Measures

This study utilized measures of fit that either evaluated global model fit or structural model fit. The structural measures of fit fall into one of three groups. The first group are measures of fit that rely on the $\triangle T_{ML}$ and $\triangle df$ between the SM_{sat} and SM_{target} models - referred to hereafter as conventional SFIs. The second group are measures of fit that take into account SM_{null} and SM_{sat}, as described by Lance et al. (2016) - the

SM-LV SFIs. The third group are measures of fit that depend on SM_{sat} in order to evaluate structural fit using $PATH_{target}$ and $PATH_{target}$, as described by Hancock and Mueller (2011) - the SM-MV SFIs.

Global Measures of Fit. Global measures of fit evaluate the model as a whole (i.e., both the measurement and structural models simultaneously). This study used both measures of exact fit (e.g., test statistics) and approximate fit (e.g., AFIs). These are introduced below.

Likelihood ratio ratio test statistic (T_{ML}). Assuming multivariate normality, this test statistic is χ^2 distributed and its expected value is equal to its degrees of freedom (Bollen, 1989).

$$\begin{aligned} \hat{T}_{ML} &= \hat{F}_{ML} * N \\ \hat{T}_{ML} &\sim \chi^2 \end{aligned} \tag{17}$$

Critical values are determined as a function of degrees of freedom and α. T_{ML} is a sample estimate of model misfit and produces a χ^2 test of exact model fit (absolute perspective) that does not adjust for model complexity. χ^2 is highly sensitive to sample size, as negligible deviations in fit will produce a statistically significant χ^2 (i.e., indicating the model does not fit the data). Hu and Bentler (1999) recommend an α of 0.05 when determining the critical value.

For a population inference on model fit, λ, the noncentrality parameter is utilized: $\lambda = \hat{T}_{ML} - df$. When λ is is divided by N - 1, McDonald's d results:

$$d = \frac{\lambda}{N-1} = \frac{\hat{T}_{ML} - df}{N-1} \tag{18}$$

McDonald's measure of centrality (Mc). Mc is interpreted from the absolute perspective, does not adjust for model complexity, and is population based. McDonald (1989) defines Mc as:

$$Mc = e^{-\frac{1}{2}d_{alt}} \tag{19}$$

Hu and Bentler (1999) recommend a cut-off of ≥ 0.90 for Mc.

Root mean squared error of approximation (RMSEA). The RMSEA is a test of close fit from the absolute perspective, adjusts for model complexity, and is population based.

This AFI is interpreted as the amount of misfit in the model per degree of freedom and is the only AFI with a known distribution (i.e., confidence intervals can be constructed (R. C. MacCallum, Browne, & Cai, 2006)). Steiger and Lind (May, 1980) define it as:

$$RMSEA = \sqrt{max(\frac{d_{alt}}{df}, 0)}, \tag{20}$$

Hu and Bentler (1999) recommend a cut-off of ≤ 0.06.

Standardized root mean square residual (SRMR). The SRMR is based on the residual correlation matrix with j rows and k columns. The equation for SRMR is given below, where p^* equals the number of unique elements in the covariance matrix, given by $\frac{p*(p+1)}{2}$. This AFI comes from the absolute perspective; it does not adjust for model complexity and is sample based. Bollen (1989) defines it as:

$$SRMR = \sqrt{\sum_j \sum_k \frac{r_{jk}^2}{p^*}} \tag{21}$$

Hu and Bentler (1999) recommend a cut-off of ≤ 0.08.

Comparative fit index (CFI). The CFI is an incremental AFI, as it relies on the estimation of a baseline model (d_{null}) and is interpreted as the improvement in fit over the baseline model. This AFI does not adjust for model complexity, and is population based. Bentler (1990) defines it as:

$$CFI = 1 - \frac{max(d_{alt}, 0)}{max(d_{alt}, d_{null}, 0)}, \tag{22}$$

Hu and Bentler (1999) recommend a cut-off of ≥ 0.95.

Tucker-Lewis Index (TLI). The TLI is an incremental fit index that adjusts for model complexity and is sample based. It scales the amount of misfit in both the alternative and baseline models as a function of their degrees of freedom. Tucker and Lewis (1973) define it as:

$$TLI = \frac{\frac{T_{null}}{df_{null}} - \frac{T_{alt}}{df_{alt}}}{\frac{T_{null}}{df_{null}} - 1} \tag{23}$$

Hu and Bentler (1999) recommend a cut-off of ≥ 0.95.

Structural Measures of Fit. Similar to the global measures of fit, this study utilized both test statistics and fit indices. All measures relied on the estimation of the

correlated factors model or SM_{sat}. The specific structural measures are grouped as conventional SFIs, SM-LV SFIs, and SM-MV SFIs.

Conventional SFIs. The conventional SFIs included the test of nested model comparison and the root mean squared error of approximation - path model (McDonald & Ho, 2002). Using the change in $\hat{T}_{ML}$ and degrees of freedom between SM_{target} and SM_{sat}: $\triangle\hat{T}$ and $\triangle df$, respectively; a test statistic results and its significance is evaluated as a function of $\triangle df$ and α - typically 0.05.

The RMSEA-P is interpreted as the amount of misfit in the structural model per structural degree of freedom. In order to estimate RMSEA-P, the structural analog of d must be estimated using the sample size: $d_p = \frac{\triangle\hat{T}_{ML} - \triangle df}{n-1}$.

$$RMSEA - P = \sqrt{max(\frac{d_P}{\triangle df}, 0)} \tag{24}$$

McDonald and Ho (2002) do not offer guidelines for RMSEA-P, however, in the literature RMSEA-P < 0.08 is typically regarded as indicating acceptable structural model fit.

SM-LV SFIs. The SFI measures from the SM-LV approach conceive the fit of the hypothesized structural model (SM_{target}) falling somewhere between the worst (SM_{null}) and the best fitting (SM_{sat}) structural models, with the measurement model remaining constant across the three models. Using $\hat{T}_{ML}$ and the degrees of freedom from all three structural models, SFIs are estimated that correspond to the distance SM_{target} is from perfect fit - C10 SFIs; as well as, the improvement over the worst fit - C9 SFIs. Lance et al. (2016) offer three methods of constructing these SFIs that are either noncentrality parameter (NCP) based, ratio-based, or neither. This study focused on NCP and ratio based forms of the SFIs. These are presented below.

C9 SFIs. The NCP SFI is interpreted as the proportion of structural model fit realized by the target model over the null structural model, in non-centrality parameter units:

$$NCP - C9 = \frac{(T_{null} - T_{target}) - (df_{null} - df_{target})}{(T_{null} - T_{sat}) - (df_{null} - df_{sat})} \tag{25}$$

The ratio based SFI is interpreted as the proportion of structural model fit realized by the target model over the null structural model, after taking into account model

complexity:

$$T : df - C9 = \frac{\frac{T_{null}}{df_{null}} - \frac{T_{target}}{df_{target}}}{\frac{T_{null}}{df_{null}} - \frac{T_{sat}}{df_{sat}}} \tag{26}$$

Lance et al. (2016) recommend a cut-off of ≥ 0.99 for both C9 SFIs.

C10 SFIs. The NCP SFI is interpreted as the proportion of structural model misfit introduced by the target model relative to the saturated structural model, in non-centrality parameter units:

$$NCP - C10 = \frac{(T_{target} - T_{sat}) - (df_{target} - df_{sat})}{(T_{null} - T_{sat}) - (df_{null} - df_{sat})}. \tag{27}$$

The ratio based SFIs is interpreted as the proportion of structural model misfit introduced by the target model relative to the saturated structural model after taking model complexity into account:

$$T : df - C10 = \frac{\frac{T_{target}}{df_{target}} - \frac{T_{sat}}{df_{sat}}}{\frac{T_{null}}{df_{null}} - \frac{T_{sat}}{df_{sat}}} \tag{28}$$

Lance et al. (2016) provide a cut-off of ≤ 0.01 for all C10 SFIs.

SM-MV SFIs. The SFI measures constructed from the SM-MV approach use the model-implied moments from SM_{sat} to inform subsequent path models: the target ($PATH_{target}$) and null ($PATH_{null}$). As a result, pseudo statistics (e.g., $\tilde{T}_{ML}$) and their respective degrees of freedom (ν) are used to estimate SFIs in the same fashion as AFIs (Hancock & Mueller, 2011; McNeish & Hancock, 2018). It then follows that $\tilde{d}$ is determined via: $\frac{\tilde{T}-\nu}{N-1}$.

Structural Mc (Mc.sfi) is a population fit index that utilizes the estimate of the non-centrality parameter and is estimated as:

$$Mc.sfi = e^{-\frac{1}{2}\tilde{d}} \tag{29}$$

Structural RMSEA (RMSEA.sfi) offers the amount of misfit in the structural model per degree of freedom. Therefore, this index is absolute in nature and adjusts for model complexity (i.e., interpreted as the degradation in fit from the saturated model).

$$RMSEA.sfi = \sqrt{max(\frac{\tilde{d}}{\nu}, 0)} \tag{30}$$

Structural SRMR (SRMR.sfi) provides an absolute judgment on structural fit to data, where $r_{S_{jk}}$ corresponds to the j^{th} and k^{th} element of the residual correlation matrix, r_S, stemming from the standardized difference between $\tilde{\mathbf{\Omega}}_{sat}$ and $\tilde{\mathbf{\Omega}}_{target}$.

$$SRMR.sfi = \sqrt{\sum_j \sum_k \frac{r_{S_{jk}}^2}{p^*}} \tag{31}$$

Structural CFI (CFI.sfi) provides an incremental judgment on structural model fit and is estimated by:

$$CFI.sfi = 1 - \frac{max(\tilde{d}_{target}, 0)}{max(\tilde{d}_{target}, \tilde{d}_{null}, 0)} \tag{32}$$

Structural TLI (TLI.sfi) provides an incremental judgment on structural model fit and is estimated by:

$$TLI.sfi = \frac{\frac{\tilde{T}_{null}}{\nu_{null}} - \frac{\tilde{T}_{target}}{\nu_{target}}}{\frac{\tilde{T}_{null}}{\nu_{null}} - 1} \tag{33}$$

4. Type I Error Simulation

Method

Monte Carlo Design. In this study, I systematically varied measurement quality, model size, and the number of observations in each group. These study conditions are detailed.

Measurement quality (MQ). The impact of measurement quality on structural model assessment was investigated in a manner similar to that of previous research, which utilized standardized factor loadings for this purpose (Hancock & Mueller, 2011; Heene et al., 2011; McNeish & Hancock, 2018; Moshagen & Auerswald, 2017). The degree of measurement quality is quantified via indicator reliability (IR). In the study, three levels of MQ were investigated: low ($\Lambda = 0.4$, IR = 0.16); moderate ($\Lambda = 0.6$, IR = 0.36); and high ($\Lambda = 0.8$, IR = 0.64).

Number of manifest variables per latent variable (p:f). In the past, researchers have used p:f to investigate the impact that model size has on AFIs (Fan & Sivo, 2009; Heene et al., 2011; Kenny et al., 2015; Lance et al., 2016). A desirable property of having a larger p:f ratio is that the resulting latent construct will be more reliably measured than when p:f is smaller. At the same time, larger p:f ratios place more of a burden on estimation, as it is tasked with reproducing a larger observed variance-covariance matrix than when p:f is smaller. Another consequence of a larger p:f ratio is the effect it has on the degrees of freedom in the variance structure; specifically, it becomes overwhelmed by the measurement model (McDonald & Ho, 2002). In this study, two levels of model size were investigated: small (p:f = 3, 3 MVs per LV) and moderate (p:f = 5, 5 MVs per LV).

Sample size. The total sample size across all simulation conditions was fixed at N = 2000. This afforded the opportunity to investigate the effect of unbalanced group sample sizes on the performance of the measures of fit. The impact of group sample size is quantified in terms of the ratio between N and the number of MVs (N:MVs); this metric is more informative than using N (Osborne & Costello, 2004) and has been used in previous studies (Fan & Sivo, 2009). In this study, three levels of subgroup sample sizes were investigated: balanced (1000 observations in each group), unbalanced with groups

1 and 2 containing 600 and 1400 observations each, respectively ($n_{g1} = 600, n_{g2} = 1400$); and unbalanced with groups 1 and 2 containing 1400 and 600 observations each, respectively ($n_{g1} = 1400, n_{g2} = 600$). The least information available occurs when group sample size is 600 and p:f is 5 (N:MV = 24), while the most information available occurs when group sample size is 1400 and p:f is 3 (N:MV = 93.33).

Table 4

Type I Error Simulation Conditions

Note. MQ = measurement quality; p:f = number of manifest variables per latent variable; G1.n = group one sample size; G2.n = group two sample size.

Condition No.	MQ	p:f	G1.n	G2.n
1	0.400	3	600	1400
2	0.600	3	600	1400
3	0.800	3	600	1400
4	0.400	5	600	1400
5	0.600	5	600	1400
6	0.800	5	600	1400
7	0.400	3	1000	1000
8	0.600	3	1000	1000
9	0.800	3	1000	1000
10	0.400	5	1000	1000
11	0.600	5	1000	1000
12	0.800	5	1000	1000
13	0.400	3	1400	600
14	0.600	3	1400	600
15	0.800	3	1400	600
16	0.400	5	1400	600
17	0.600	5	1400	600
18	0.800	5	1400	600

These study conditions were fully crossed, resulting in 18 unique simulation conditions [3 (MQ) * 2 (p:f) * 3 (n_{group})] from which to examine the sampling distribution of the structural measures of fit, see Table 4 for each unique combination. A total of 1000 data sets were generated for each condition and ultimately analyzed.

Outcomes. After estimation of the latent variable and path models, the measures of fit were computed for each replication across all simulation conditions.

Univariate statistics. Using the psych package (Revelle, 2018), summary statistics were estimated to determine the expected value and the degree to which this value varied across simulation conditions. Aside from these moments, skew and kurtosis estimates were investigated to determine the degree to which the distribution of the fit measures deviated from normality.

Effect of design factors. In order to determine the effect of measurement quality, model size, and sample size (balanced versus unbalanced group size) on the measures of fit, an analysis of variance (ANOVA) was performed in R using the stat package (R Core Team, 2017). Therefore, a series of ANOVAs (one for each measure of fit) were performed with all main (e.g., MQ) and interaction effects (e.g., MQ * p:f) included in the model. Due to the large number of observations (1000 replications per condition), statistical significance of the main and interaction effects were not considered. Instead, an effect size, partial η^2, was used to assess the impact of design factors. Partial η^2 was estimated using the lsr package (Navarro, 2015) in R. Specifically, the guidelines proposed by Cohen (1988) were utilized where estimates of 0.01, 0.06, and 0.14 correspond to small, medium, and large effect sizes.

Empirical cut-offs. Upon estimating the ANOVAs, if a given measure of fit is found to be affected by design factors as evidenced by partial $\eta^2 \geq 0.14$, then it was no longer considered. For the measures still being considered, their values at the 95^{th} and 99^{th} percentiles (collapsed over all simulation design factors) were recorded.

Due to the estimation model being correct in the population, all measures of fit should indicate a close fitting model, regardless of design factors. This is because the reliability paradox only presents itself if the structural model is misspecified (Heene et al., 2011;

Moshagen & Auerswald, 2017).

It was hypothesized that:

- The AFIs, RMSEA-P, and the SM-MV SFIs from the absolute perspective should be near 0.0 - with the exception of Mc, while those from the incremental perspective and Mc should be near 1.0. With respect to the test statistics, their values should approach their degrees of freedom. With respect to the SM-LV SFIs, the C9 SFIs should approach 1.0 and the C10 SFIs should approach 0.0.
- A large effect (i.e., partial $\eta^2 \geq 0.14$) would result for model complexity since the expected value for T_{ML} is its degrees of freedom. When more indicators are included, the model's degrees of freedom will increase.
- All measures of fit that are determined using a latent variable model would have negligible partial η^2 (e.g., AFIs, RMSEA-P, and SM-LV SFIs) for all study factors.
- Measurement quality would have an adverse affect on the SM-MV SFIs due to these SFIs depending on the quality of the latent variables. As measurement quality decreases, the SM-MV SFI standard deviations should increase, thus making them highly unreliable for assessing structural model misfit.

Results

Following the execution of the simulation, a convergence rate of 100 percent was achieved. This afforded the opportunity to fully examine the sampling distribution of the various measures of fit. As such, a balanced examination of study conditions was carried out because there were an equal number of converged replications per simulation condition. Thus, the resulting partial η^2 estimates reflect true proportion of variance explained. In sum, the results reported below are based on 18,000 observations.

Sampling Distribution. Measures of central tendency and dispersion were estimated for each measure of fit. Descriptive statistics are presented together below depending on whether they assess global model fit, structural model fit via conventional methods, SM-LV, or SM-MV.

Global Measures. The expected values for the test statistic (χ^2), as well as the incremental and absolute AFIs, were as hypothesized. The mean χ^2 value was 400.916 with 397 degrees of freedom on average, indicating the estimation model is correct in the population. In terms of incremental AFIs, the estimates for the CFI and TLI were near 1.000: 0.997 and 0.999, respectively; with near zero standard deviations. In terms of absolute AFIs, the estimate for Mc was 0.999 on average and did not fall below 0.968; whereas, the RMSEA and the SRMR had mean estimates of 0.004 and 0.026, respectively, and neither had estimates greater than 0.035.

See Table 5 for all relevant statistics for these measures of fit.

SM - Conventional. These measures of fit on average were found to indicate a well fitting model. Specifically, the $\triangle df$ between SM_{sat} and SM_{target} was 22 across all replications. The expected value for χ^2 was 22.251, demonstrating that on average the $\triangle\chi^2$ correctly detects no structural model misfit. With respect to the RMSEA-P, its estimate was 0.005 (SD = 0.006), on average; therefore, no structural model misfit was detected.

See Table 6 for all relevant statistics for these measures of fit.

SM - Latent Variable. The C9 and C10 SFIs behaved as expected regardless of how they were estimated (i.e., non-centrality or ratio based). Specifically, the mean estimate for the C10 SFIs was 0.0001 and the mean estimate for the C9 SFIs was 1.000, on average. The standard deviation of these SFIs was 0.003.

See Table 7.

SM - Manifest Variable. The SFIs that stem from the approach of Hancock and Mueller (2011) objectively performed poorly. Recall that the data generating model and the estimation model are identical. Further, recall that the degrees of freedom is the expected value for χ^2 and the two should be in close agreement. Across all 18,000 replication, the degrees of freedom was 22 (SD = 0); therefore, the critical value is 33.92 or 40.29 when α is 0.05 and 0.01, respectively. The mean estimate of χ^2 was 112.60 ($\sigma = 349.49$) across all replications. It is troubling that over 75 percent of the replications had χ^2 values greater than 34.35 and 50 percent were 57.14 or greater.

Therefore, an overwhelming majority of the time, the pseudo χ^2 test statistic incorrectly identified a structural model that did not fit in the population.

With respect to the SFIs from the incremental perspective, the mean estimates for CFI and TLI were 0.987 ($\sigma = 0.023$) and 0.979 ($\sigma = 0.037$), respectively. In terms of those from the absolute perspective, the mean estimate for Mc was 0.979 ($\sigma = 0.036$); whereas the mean estimates for RMSEA and SRMR were 0.05 ($\sigma = 0.04$) and 0.04 ($\sigma = 0.02$), respectively. Although these SFIs, on average, approach their hypothesized values (e.g., 0.0 for RMSEA and 1.0 for CFI), the range of their values was alarming. Specifically, the maximum value observed for the RMSEA was 0.83 and the minimum value observed for the CFI was 0.295.

See Table 8 for all relevant statistics for these measures of fit. The distributions of the SM-MV SFIs versus their counterparts global counterparts are plotted together (see Figures 2, 3, and 4).

Moving forward, the conventional approach for assessing structural model misfit and that of Lance et al. (2016) will be reported together.

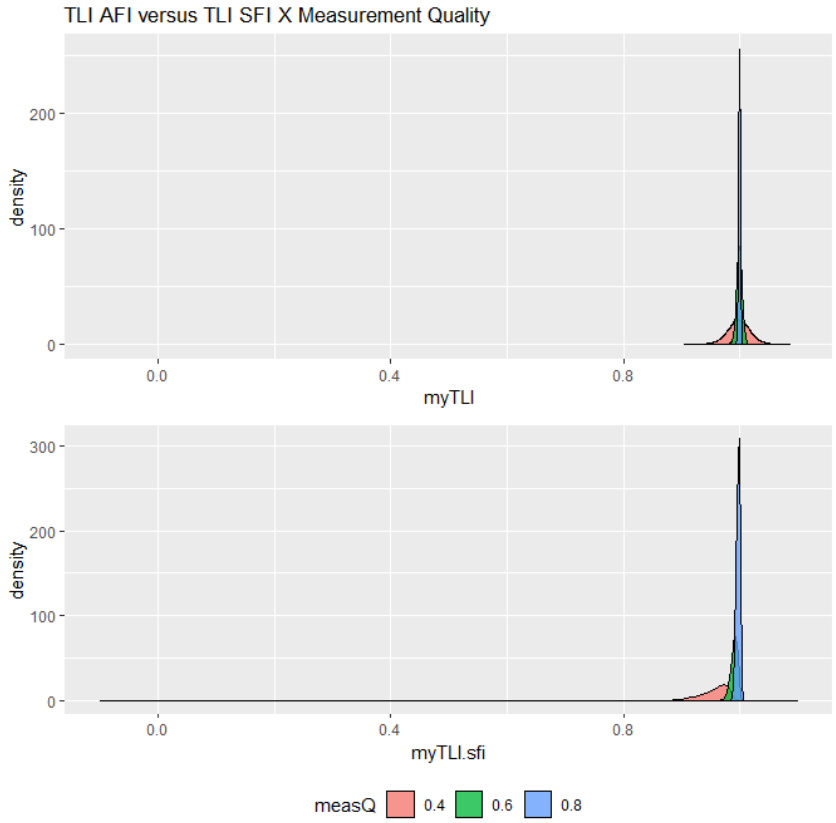

Figure 2. TLI Comparison

Global TLI distribution versus SM-MV TLI distribution. Note. myTLI utilizes the manually specified baseline model; *.sfi = structural measure of fit.

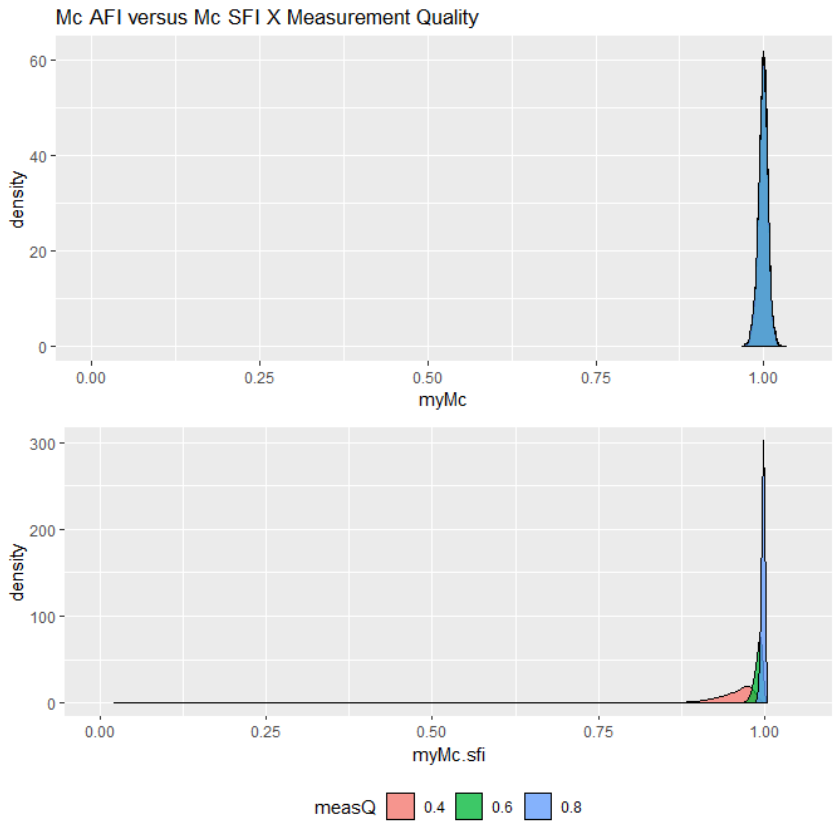

Figure 3. Mc Comparison

Global Mc distribution versus SM-MV Mc distribution. Note. myMc is manually estimated; *.sfi = structural measure of fit.

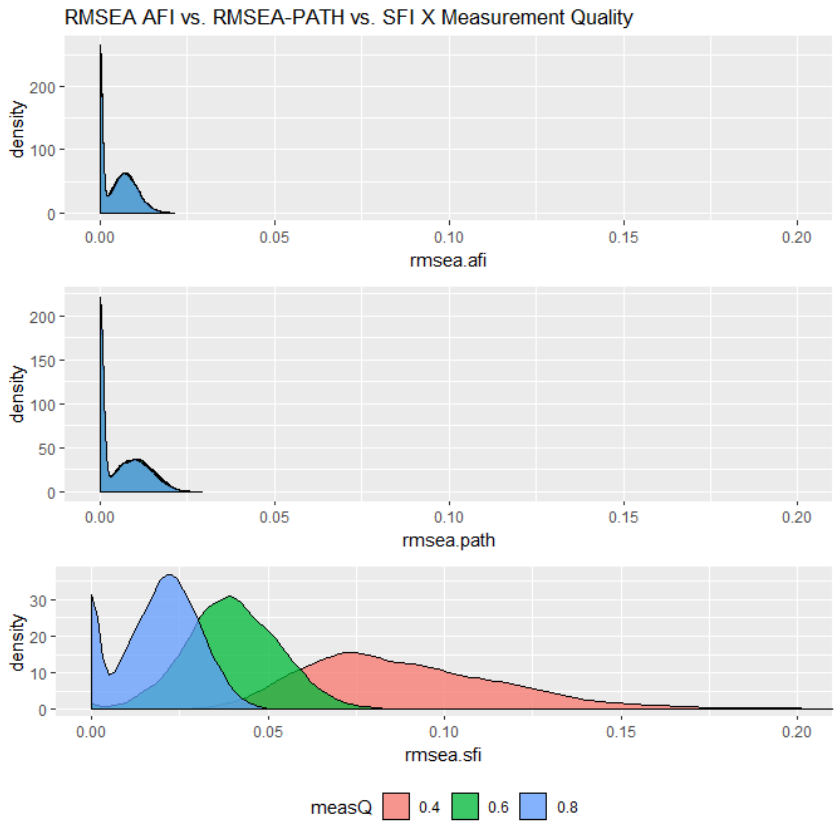

Figure 4. RMSEA Comparison

Global RMSEA versus RMSEA-P versus SM-MV RMSEA Distributions. *.sfi = structural measure of fit.

Table 5

Descriptive Statistics - Global Measures of Fit

*Note. myCFI and myTLI utilize the manually specified baseline model; *.afi = global measure of fit; N = total number of replications*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
target.csq	18,000	400.916	199.452	138.951	202.813	379.178	597.989	723.934
target.df	18,000	397.000	195.005	202	202	397	592	592
target.fmin	18,000	0.200	0.100	0.069	0.101	0.190	0.299	0.362
target.d	18,000	0.002	0.014	−0.063	−0.007	0.001	0.011	0.066
myCFI	18,000	0.997	0.007	0.920	0.997	0.999	1	1
myTLI	18,000	0.999	0.012	0.905	0.997	1.000	1.002	1.085
myMc	18,000	0.999	0.007	0.968	0.995	0.999	1.004	1.032
rmsea.afi	18,000	0.004	0.004	0	0	0.003	0.01	0
srmr.afi	18,000	0.026	0.002	0.017	0.024	0.026	0.027	0.035

Table 6

Descriptive Statistics: Conventional Measures

Note. delta.csq = $\triangle\chi^2$; delta.df = $\triangle df$; rmsea.path = RMSEA-P; N = total number of replications

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
delta.csq	18,000	22.251	6.718	5.199	17.417	21.586	26.340	59.281
delta.df	18,000	22.000	0.000	22	22	22	22	22
path.d	18,000	0.0001	0.003	−0.008	−0.002	−0.0002	0.002	0.019
rmsea.path	18,000	0.005	0.006	0	0	0	0.01	0.029

Table 7

Descriptive Statistics: SM-LV Measures

Note. $smlv.sat.csq = \chi^2 : SM_{sat}$; $smlv.sat.df = df : SM_{sat}$; $smlv.null.csq = \chi^2 : SM_{null}$; $smlv.null.df = df : SM_{null}$ N = *total number of replications*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
smlv.sat.csq	18,000	378.665	199.459	12.658	180.616	357.950	575.787	697.118
smlv.sat.df	18,000	375.000	195.005	180	180	375	570	570
smlv.null.csq	18,000	3,412.728	1,348.307	1,317.390	2,128.107	3,390.338	4,733.238	5,677.888
smlv.null.df	18,000	410.000	195.005	215	215	410	605	605
C9.ncp	18,000	1.000	0.003	0.973	0.998	1.000	1.002	1.013
C9.perDF	18,000	1.000	0.003	0.969	0.998	1.000	1.002	1.017
C10.ncp	18,000	0.0001	0.003	−0.013	−0.002	−0.0001	0.002	0.027
C10.perDF	18,000	0.0001	0.003	−0.017	−0.002	−0.0002	0.002	0.031

Table 8

Descriptive Statistics: SM-MV Measures

*Note. myCFI and myTLI utilize the manually specified baseline model; *.sfi = structural measure of fit; target.csq.sfi =* $\chi^2 : PATH_{target}$*; target.df.sfi = df :* $PATH_{target}$*; target.fmin.sfi =* $F_{ML} : PATH_{target}$*; N = total number of replications*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
target.csq.sfi	18,000	112.596	349.488	6.852	34.349	57.142	128.209	15,293.210
target.df.sfi	18,000	22.000	0.000	22	22	22	22	22
target.fmin.sfi	18,000	0.056	0.175	0.003	0.017	0.029	0.064	7.647
myCFI.sfi	18,000	0.987	0.023	0.295	0.983	0.994	0.998	1.000
myTLI.sfi	18,000	0.979	0.037	−0.122	0.973	0.991	0.997	1.004
myMc.sfi	18,000	0.979	0.036	0.022	0.974	0.991	0.997	1.004
rmsea.sfi	18,000	0.050	0.040	0.000	0.024	0.040	0.069	0.833
srmr.sfi	18,000	0.041	0.017	0.012	0.028	0.036	0.049	0.153

Effect of Simulation Conditions. To determine the effect model size (pF, Factor A), measurement quality (MQ, Factor B), and balanced group sample sizes (bal.n, Factor C) on each measures of fit, an analysis of variance was performed. All possible interaction effects were investigated, along with their respective main effects.

Global Measures. As expected, model size was an influential factor in the variability of χ^2 (pF, partial $\eta^2 = 0.98$); whereas, McDonald's d was unaffected by model size or any other factors due to its being based on the non-centrality parameter.

Table 9

Effect of Design Factors: Global Measures

*Partial $\eta^2 < 0.01$ are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); bal.n = balanced groups; myCFI and myTLI utilize the manually specified baseline model; *.afi = global measure of fit; N = total number of replications.*

	pF(A)	MQ(B)	bal.n(C)	AxB
target.csq	0.980			
target.fmin	0.980			
target.d				
myCFI		0.181		
myTLI				
myMc				
rmsea.afi				
srmr.afi	0.096	0.342	0.063	0.029

With respect to the incremental AFIs, 18 percent of the variability in CFI across replications was explained by measurement quality (partial $\eta^2 = 0.18$), corresponding to a large effect; however, no factor explained the variability in TLI estimates across the simulation conditions. In terms of absolute AFIs, the Mc was found to be unaffected by study conditions as was the RMSEA. On the other hand, the interaction between model size and measurement quality (pF*MQ, partial $\eta^2 = 0.03$) was found to have a small to medium effect on SRMR estimates, with the lion's share stemming from the main effect

of measurement quality (MQ partial $\eta^2 = 0.34$). See Table 9 for all partial η^2 estimates. Based on the sensitivity of SRMR to model size, measurement quality and the interaction of these factors, the SRMR was no longer considered. Likewise, CFI was no longer considered because it was largely dependent on the level of measurement quality. On the other hand, Mc and RMSEA from the absolute perspective were found to be unaffected by study conditions. In terms of incremental AFIs, the TLI was found to be unaffected by study conditions, and its variability across simulation conditions was attributed to sampling error. Therefore, the RMSEA, Mc, and the TLI were retained, and cut-off values were determined for them at their respective 99^{th} and 95^{th} percentiles.

SM - Conventional and Latent Variable. It was found that neither the Lance et al. (2016) SFIs, nor the RMSEA-P, nor the $\triangle\chi^2$ approaches were affected by conditions manipulated in this study. By taking the difference of the χ^2 between the target and the saturated model, the effect of model complexity found earlier is controlled for. Based on the performance of the RMSEA-P, the C9 indices, and the C10 indices, empirical cut-offs were established for them corresponding to their respective 99^{th} and 95^{th} percentiles.

SM - Manifest Variable. As stated earlier, SFIs from the SM-MV approach behaved in a concerning way. This behavior can be better understood when consulting Table 10. Specifically, the interaction effect of model size and measurement quality had a small effect on the pseudo test statistic, $\tilde{\chi}^2$ (partial $\eta^2 = 0.02$), with the main effect of measurement quality corresponding to a medium effect (partial $\eta^2 = 0.075$). The interaction of model size and measurement quality had a medium-to-large effect on the CFI, TLI, Mc, RMSEA, and SRMR with partial η^2 ranging from 0.10 to 0.16. Across all measures, the effect of measurement quality was alarming: partial η^2 was 0.36 for both the CFI and TLI; 0.39 for Mc; and 0.65 and 0.70 for the RMSEA and SRMR, respectively. In terms of model size, medium-to-large effects were found and partial η^2 estimates ranged from 0.085 to 0.22.

In light of the performance of the SM-MV SFIs, it was decided to no longer consider them in the subsequent Power simulation. This decision was made in response to the

Table 10

Effect of Design Factors: SM-MV Measures

*Partial $\eta^2 < 0.01$ are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); bal.n = balanced groups; myCFI and myTLI utilize the manually specified baseline model; *.sfi = structural measure of fit; N = total number of replications.*

	pF(A)	MQ(B)	bal.n(C)	AxB
target.csq.sfi	0.015	0.075		0.020
target.fmin.sfi	0.015	0.075		0.020
target.d.sfi	0.015	0.075		0.020
myCFI.sfi	0.085	0.363		0.100
myTLI.sfi	0.085	0.364		0.099
myMc.sfi	0.095	0.388		0.113
rmsea.sfi	0.175	0.646		0.104
srmr.sfi	0.219	0.704	0.027	0.161

medium-to-large effect that the interaction of model size and measurement quality had on these measures. It is infeasible to determine a single cut-off value for these measures that is appropriate over varying modeling conditions.

Empirical Cut-Offs. For measures of fit where larger values indicate a better fitting model (e.g., Mc and TLI), cut-offs that correspond to an $\alpha = 0.05$ (95^{th} percentile) and $\alpha = 0.01$ (99^{th} percentile) correspond to the value at which 5% or 1% of the estimated values over all simulation replications are at or below this value. For measures of fit where smaller values indicate a better fitting model, $\alpha = 0.05$ (95^{th} percentile) and $\alpha = 0.01$ (99^{th} percentile) correspond to the value where 5% or 1% of the replications are at or above their respective values. Table 11 contains all cut-off values for each the measures of fit to be utilized in the Power simulation. In an effort to communicate the value needed for each fit measure that represents near certainty that the model fits the data, the value at the 99.9^{th} percentile also was determined and also is presented in Table 11.

Table 11

Empirically Derived Cut-Off values

Note. 95, $\alpha = 0.05$*; 99,* $\alpha = 0.01$*; 99.9,* $\alpha = 0.001$

*myCFI and myTLI utilize the manually specified baseline model; myMc = McDonald's measure of centrality; *.afi = global measure of fit; perDF =* $\frac{T_{ML}}{df}$*; ncp = non-centrality parameter; rmsea.path = RMSEA-P*

Measure	95	99	99.99
myTLI	0.979	0.960	0.939
myMc	0.987	0.981	0.973
rmsea.afi	0.012	0.015	0.018
C9.ncp	0.994	0.989	0.983
C9.perDF	0.994	0.989	0.982
C10.ncp	0.006	0.011	0.017
C10.perDF	0.006	0.011	0.018
rmsea.path	0.017	0.021	0.025

The empirical cumulative distribution functions of the global fit indices are plotted among each other (Figure 5). In this plot, the vertical line corresponds to the cut-off values presented above, while the blue line corresponds to the critical value with an α of 0.01 and the red line corresponds to the critical value with an α of 0.05. As a point of reference, the value that corresponds to the 99.99^{th} percentile is represented by the dashed black line. In a similar fashion, the empirically based critical values for the structural measures of fit are plotted (Figure 6). It is interesting to note that regardless of the method from which the C9 and C10 SFIs are estimated (e.g., non-centrality based), their critical values are the same. For instance, at the 99^{th} percentile, this value is 0.989 and 0.011 for the C9 and C10 SFIs, respectively; whereas, at the 95^{th} percentile, the critical values are 0.994 and 0.006 for the C9 and C10 SFIs, respectively. With respect to the RMSEA-P, its critical values are larger than those for the C10 SFIs, with a value of 0.021 and 0.017 corresponding to the 99^{th} and the 95^{th} percentiles, respectively.

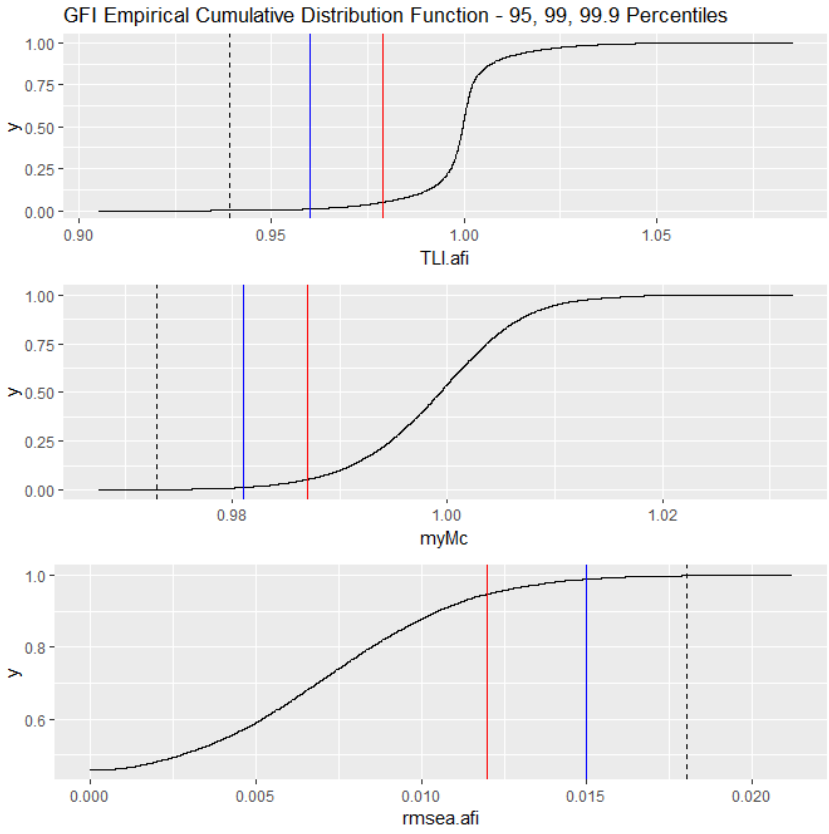

Figure 5. ECDF: Global Measures

Note. TLI.afi = Tucker-Lewis index; myMc = McDonald's measure of centrality

Summary of Results

As hypothesized, χ^2 was impacted by model size, while McDonald's d was not. The global fit index CFI was impacted by measurement quality and replicates previous research (Hancock & Mueller, 2011; Kang et al., 2016). Interestingly, a non-negligible effect of measurement quality on SRMR was found; this was not the case for RMSEA. With respect to the Conventional and SM-LV measures of fit, all were unaffected by the study designs, and their variability was solely attributed to sampling error. On the other hand, the SM-MV measures of fit were impacted greatly by measurement quality and produced a wide range of estimates. The effect of measurement quality on the SM-MV SFIs illustrates the importance of estimating the measurement and structural model simultaneously, rather than via a two-stage estimation process.

The SM-MV approach relies on the LV model's ability to account for measurement

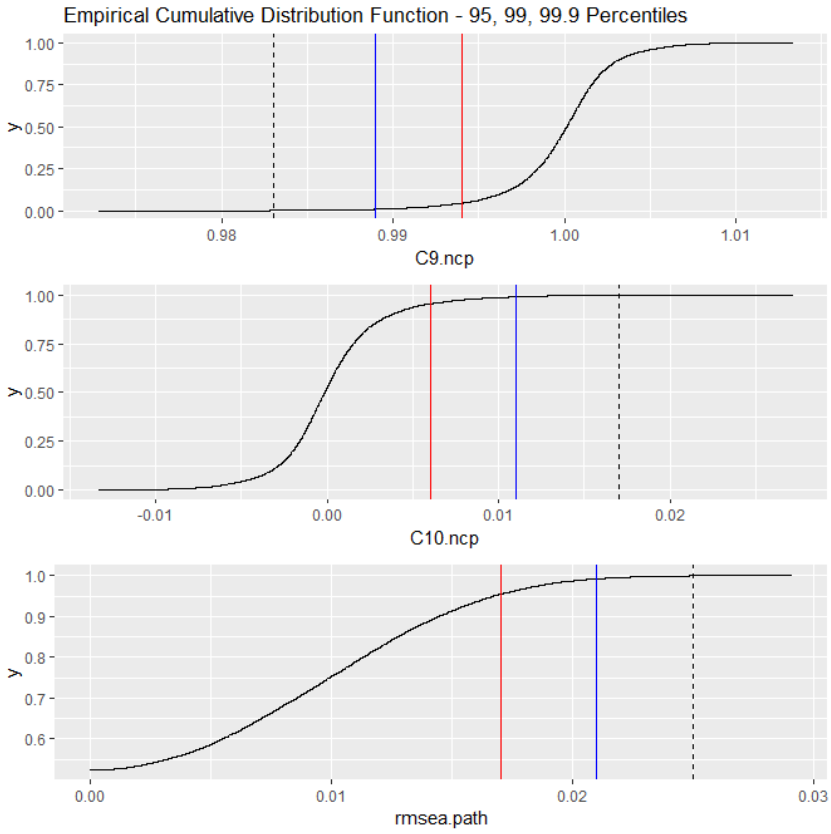

Figure 6. ECDF: Structural Measures

ncp = non-centrality parameter; rmsea.path = RMSEA-P

error. As measurement quality decreases, so does the proportion of true score variance within its respective LV. A consequence of this is that the relationships among the LVs will be overestimated. Therefore, when $\hat{\Phi}$ is inserted into the subsequent path model a tremendous task is asked of its estimator. Specifically, it is charged with reproducing the target latent variance-covariance matrix without taking into account the uncertainty in $\hat{\Phi}$. For this reason, when the population model was estimated in the path analysis, the SM-MV SFIs were unable to detect that the estimation model was correct and overwhelmingly rejected the model.

In conclusion, this simulation provides evidence for approaches that result from single-stage estimation. With respect to structural measures, the RMSEA-P, C9 and C10, and $\triangle\chi^2$ were found to be unaffected by design factors. With respect to global measures, the RMSEA, Mc, χ^2, and TLI were found to be unaffected by design factors.

5. Power Simulation

The cut-offs retained from the Type I Error simulation served as a means to evaluate the relative performance of the SFI measures of fit in their ability to detect structural misspecifications and whether their relative performance varied as a function of type and/or severity of the structural misspecification.

Method

Monte Carlo Design. As in the Type I Error simulation, I systematically varied measurement quality, model size, and group sample sizes. These design factors were manipulated in an identical fashion as earlier. The focus of this simulation was to examine whether the type of structural model misfit or the severity of misfit impacted the structural measures. Several design factors were manipulated and are described below.

Type of structural misspecification. Group differences were generated on a single mean parameter, a single covariance parameter, and both the mean and covariance parameters simultaneously. When the mean structure was misspecified, group differences were generated on the latent mean for X3 ($\alpha_{1,3}$). When the covariance structure was misspecified, group differences were generated on the latent regression of $X2 \rightarrow Y2$ ($\beta_{5,2}$); this parameter has been utilized in previous research and was selected in an effort to remain consistent (Lance et al., 2016; McNeish & Hancock, 2018). When the mean and covariance structures were simultaneously misspecified, $\alpha_{1,3}$ and $\beta_{5,2}$ again were utilized.

Severity of structural misspecification. The group differences on a given parameter were understood as standardized differences between the two groups. In total, three levels of severity were investigated and corresponded to small, medium, and large differences between groups. Specifically, when the the mean structure was misspecified, standardized mean differences were d = 0.2 (small effect), d = 0.5 (medium effect), or d = 0.8 (large effect), and were chosen based on their being used in prior research (Fan & Sivo, 2009). For each of the three mean misspecifications, the mean of X3 for the

reference group was held at zero.

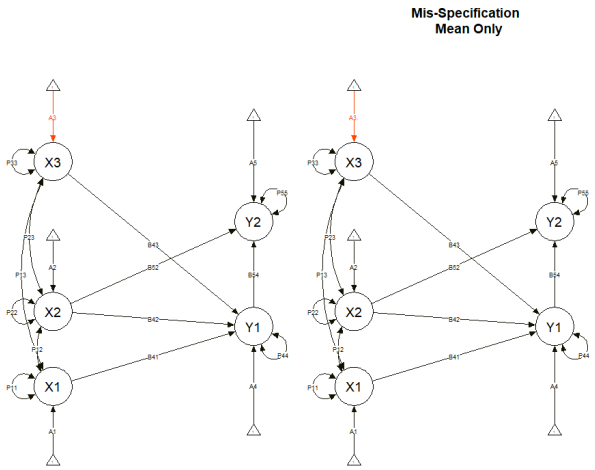

Figure 7. Path Diagram: Mean Misspecification

Note. Multiple group model with mean misspecification - in red. Exogenous latent variables: X1, X2, X3. Endogenous latent variables: Y1 and Y2.

When only the covariance structure was misspecified, the standardized differences utilized for the regression parameter were d = 0.2 (small effect), d = 0.4 (medium effect), or d = 0.6 (large effect); and these have been used in prior research (Kang et al., 2016). Depending on the severity of the covariance misspecification the population value for $B_{5,2}$ differed for each group across these conditions. See Data Generation below. When both the mean and covariance structures were simultaneously misspecified, this represented another degree of model misspecification and contained all possible combinations of mean and covariance misspecifications (i.e., $\triangle\alpha_{1,3} = 0.8$ and $\triangle\beta_{5,2} = 0.6$), resulting in 9 unique misspecification conditions.

In total, there were 15 [3 (mean) + 3 (covariance) + 9 (mean*covariance)] levels of structural model misspecifications. These misspecifications were then crossed with the varying levels of model size (p:f), measurement quality (MQ), and group sample sizes. In sum, 270 unique conditions were examined in this Monte Carlo simulation. For each condition, 1000 data sets were generated to fully assess the relative performance of the RMSEA, Mc, TLI, C9 and C10, RMSEA-P, and $\triangle\chi^2$ in detecting structural model

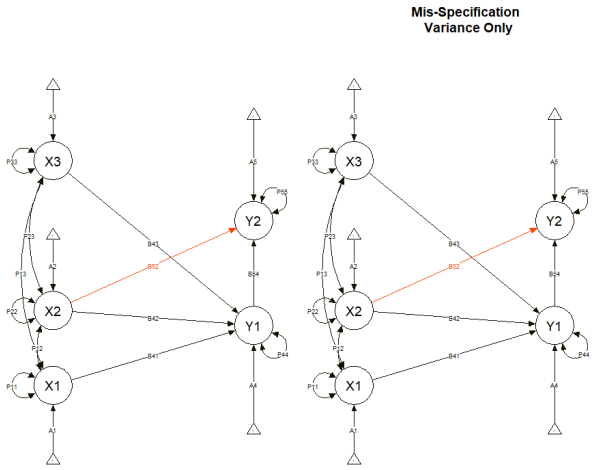

Figure 8. Path Diagram: Covariance Misspecification

Note. Multiple group model with covariance misspecification - in red. Exogenous latent variables: X1, X2, X3. Endogenous latent variables: Y1 and Y2.

misfit.

Data Generation. Data was generated using the same routine as described in Chapter 3; however, due to introducing group differences on mean and/or covariance parameters, population model matrices differed between groups.

Covariance Structure. Using Equation 15, $\mathbf{\Psi}_{5,5}$ was solved for each group based on their respective regression weights. In Table 12, the population values for $X2 \rightarrow Y2$ for the two groups are listed, as are the corresponding population values for the proportion of variance unexplained in Y2 by their respective system of equations Ψ_{55}.

Mean structure. Using Equation 16, $\nu_{1,4}$ was determined for Group 2 based on the difference on $\alpha_{1,3}$. Table 13 provides the population values for the latent intercepts for Group 1 and Group 2 due to group differences on the latent mean of X3.

Measures. The sole measure utilized in this simulation was power. Statistical power is a function of Type II error (or β) and is represented by $1 - \beta$. In the context of model fit, Type II errors occur when a measure of fit fails to detect the misfit and fails to reject the misspecified model. Therefore, statistical power of fit indices represents its probability of rejecting a poor fitting model when it is truly misspecified. Therefore,

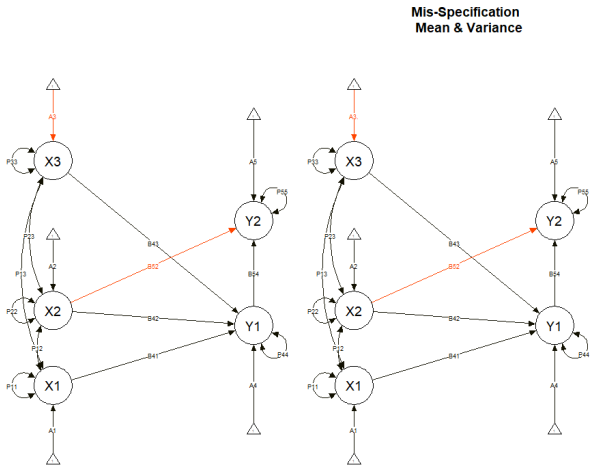

Figure 9. Path Diagram: Simultaneous Misspecification

Note. Multiple group model with simultaneous misspecifications - in red. Exogenous latent variables: X1, X2, X3. Endogenous latent variables: Y1 and Y2.

Table 12

Population Values: Covariance Structure X Group

Note. G1 = Group 1 and G2 = Group 2; B_{52} *= standardized regression weight* $X2 \rightarrow Y2$*;* Ψ_{55} *= Y2 disturbance*

	Small (d = 0.2)	Medium (d = 0.4)	Large (d = 0.6)
G1:B_{52}	0.300	0.200	0
G1:Ψ_{55}	0.744	0.820	0.910
G2:B_{52}	0.500	0.600	0.600
G2:Ψ_{55}	0.534	0.399	0.399

after estimating all models presented in Chapter 3, the measures of fit were determined and subsequently utilized to generate a hit rate representing whether the model was correctly rejected when consulting specific criteria or cut-offs and are detailed below. The hit rate is a function of *true positives* (TP) and *false negatives*. See Equation 34.

$$HitRate = \frac{TP}{TP + FN} \tag{34}$$

Global measures. The χ^2 test statistic was evaluated using the critical value that

Table 13

Population Values: Mean Structure X Group

Note. G1 = Group 1 and G2 = Group 2; ν_{13} *= latent mean for X3;* ν_{14} *= latent intercept for Y1;* ν_{15} *= latent intercept for Y2*

	ν_{11}	ν_{12}	ν_{13}	ν_{14}	ν_{15}
G1	0	0	0	1.500	0.750
G2 (small)	0	0	0.200	1.440	0.750
G2 (medium)	0	0	0.500	1.350	0.750
G2 (large)	0	0	0.800	1.260	0.750

corresponded to an α of 0.01 and 0.05. When p:f was 3, the target model's degrees of freedom was 202 and its corresponding critical values were $\alpha_{0.05} = 236.158$ and $\alpha_{0.01} = 251.677$. When p:f was 5, the model's degrees of freedom was 592 and its corresponding critical values were $\alpha_{0.05} = 649.712$ and $\alpha_{0.01} = 674.976$.

The remaining global measures of fit utilized in this simulation were Mc, TLI, and RMSEA. For each of these measures of fit, two types of cut-offs were utilized: those proposed by Hu and Bentler (1999) (HB) and the empirically determined values from the previous simulation. The TLI was judged based on values ≥ 0.95 based on HB, ≥ 0.979 for $\alpha_{0.05}$, and ≥ 0.960 for $\alpha_{0.01}$. The Mc was judged based on values ≥ 0.90 based on HB, ≥ 0.987 for $\alpha_{0.05}$, and ≥ 0.981 for $\alpha_{0.01}$. The RMSEA was judged based on values ≤ 0.06 based on HB, ≤ 0.012 for $\alpha_{0.05}$, and ≤ 0.015 for $\alpha_{0.01}$.

Structural measures. With respect to structural measures of fit, the $\triangle\chi^2$, RMSEA-P, and the ratio and non-centrality based C9 and C10 SFIs were examined. For the $\triangle\chi^2$, I utilized the critical value that corresponds to an error rate of 0.01 and 0.05. With a change in degrees of freedom of 22, the critical value for $\alpha = 0.05$ is 33.924 and for $\alpha = 0.01$ is 40.289.

For the RMSEA-P, I used a cut-off of < 0.08 as this is used in the literature, 0.017 for $\alpha = 0.05$, and 0.021 for $\alpha = 0.01$. With respect to the C9 SFIs, I used the prescribed cut-off of 0.99 as proposed by Lance et al. (2016), 0.994 for $\alpha = 0.05$, and 0.989 for $\alpha = 0.01$. With respect to the C10 SFIs, I used the prescribed cut-off of 0.01 proposed

by Lance et al. (2016), 0.006 for $\alpha = 0.05$, and 0.011 for $\alpha = 0.01$.

Outcomes. To answer research question 2, regarding differential performance of the various measures of fit based on the type of structural misspecification, only the conditions in which either the mean or the covariance structure was misspecified were considered. Therefore, following the creation of the hit rates, descriptive statistics were estimated to determine the rate at which each measure of fit was able to detect the structural model misfit. Specifically, overall descriptives statistics were generated and then descriptive statistics were computed by the type of structural model misspecification (e.g., mean only and covariance only). Afterwards, univariate ANOVAs were estimated to investigate the impact of model size (pF, Factor A), measurement quality (MQ, Factor B), unbalanced group sizes (Unbal.n, Factor C), type of misspecification (typeMis, Factor D), and severity of misspecification (severity, Factor E), as well as all possible interaction effects among these between subject factors. As in Chapter 4, partial η^2 estimates were utilized to determine what unique combination of design factors contributed to the largest proportion of variance explained in model misfit detection. In the event a given design factor was found to have a negligible effect, it was removed from the subsequent model. To answer research question 3, the same steps were taken and they only utilized conditions in which the mean and covariances structures were misspecified simultaneously.

Table 14

Taxonomy of Measures

Note. Type of evaluation X type of judgment

Dimensions	Test Statistic	Incremental	Absolute
Structural	$\triangle\chi^2$	SM-LV	RMSEA-P
Global	χ^2	TLI	RMSEA

To investigate the performance of the global measures versus the structural measures, an analysis of covariance (ANCOVA) was estimated for each type of misspecification (i.e., mean only, covariance only, and simultaneous misspecification). In each of the

ANCOVAs, the hit rate was entered as the dependent variable with the between-measure factors being global versus structural and incremental versus absolute versus test statistic. Table 14 shows the breakdown of the measures of fit based on these two dimensions. The discrepancy fit value ($\hat{F}_{ML}$) was inserted into the model as the covariate to determine what effect study design factors (e.g., model size and measurement quality) had on power rates.

Results

The convergence rate for all simulation conditions was 100 percent across all replications. As a result, a balanced investigation of factors that contribute to the performance of the measures of fit and their ability to detect structural model misfit was possible. The results below are organized by research question to allow for a more coherent presentation of the findings.

Performance With Either a Mean or Covariance Misspecification. Table 15 contains the descriptive statistics for all of the measures of fit when only the mean structure was misspecified, while Table 16 contains this information when the covariance structure was misspecified. When the covariance structure was misspecified, measures of fit indicated better model fit than when only the mean structure was misspecified. Further, the variability of the fit measures was noticeably smaller when the covariance structure was misspecified. For instance, the $\triangle\chi^2$, when it was a mean misspecification, had a mean of 112.5 (SD = 83.8), compared with 82.8 (SD = 66.5) when it was a covariance misspecification.

Impact of design factors. Of interest was to determine the impact of the manipulated design factors had on the resulting fit measure values and whether any pattern emerged across the different measures of fit based on the type of misspecification. As in the Type I Error simulation, a factorial analysis of variance was performed for each measure of fit. The between-subject factors were: model size (pF, A), measurement quality (MQ, B), reference group sample size (ref.n, C), and the severity of the misspecification (severity, D). All possible interactions were included in this model. Factorial ANOVAs

Table 15

Descriptive Statistics: Mean Misspecification

*Note. ncp = non-centrality parameter; perDF = $\frac{T_{ML}}{df}$; delta.csq = $\triangle\chi^2$; rmsea.path = RMSEA-P; myMc = McDonald's measure of centrality; myTLI utilizes the manually specified baseline model; target.csq = χ^2; *.afi = global measure of fit*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
C9.ncp	54,000	0.969	0.025	0.866	0.948	0.993	1.011
C9.perDF	54,000	0.967	0.027	0.857	0.945	0.993	1.014
C10.ncp	54,000	0.031	0.025	−0.011	0.007	0.052	0.134
C10.perDF	54,000	0.033	0.027	−0.014	0.007	0.055	0.143
delta.csq	54,000	112.508	83.755	7.239	40.591	156.124	425.039
rmsea.path	54,000	0.040	0.022	0.000	0.021	0.055	0.096
myMc	54,000	0.977	0.021	0.894	0.966	0.994	1.019
myTLI	54,000	0.971	0.035	0.770	0.956	0.994	1.082
target.csq	54,000	308.613	111.645	140.708	227.356	356.794	889.202
rmsea.afi	54,000	0.018	0.011	0.000	0.011	0.026	0.047

were performed for each type of misspecification and are presented by type of measure (e.g., global or structural). The partial η^2 estimates from these factorial ANOVAs can be found in Appendix A; however, I highlight the key findings here.

With respect to global fit measures in the context of a mean misspecification, the severity of misspecification was found to have the largest impact. The η^2 ranged from 0.67 (TLI) to 0.88 (χ^2). The main effect of measurement quality also was considerable, with η^2 ranging from 0.24 (χ^2) to 0.57 (TLI). Interestingly, reference group sample size had no impact on variability and model size had a negligible effect (less than 0.01) for all global measures except χ^2, which was expected. Several 2-way interactions were found to have an effect ($\eta^2 \geq 0.06$) and differed by global measure of fit. For the TLI,

Table 16

Descriptive Statistics: Covariance Misspecification

*Note. ncp = non-centrality parameter; perDF = $\frac{T_{ML}}{df}$; delta.csq = $\triangle\chi^2$; rmsea.path = RMSEA-P; myMc = McDonald's measure of centrality; myTLI utilizes the manually specified baseline model; target.csq = χ^2; *.afi = global measure of fit*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
C9.ncp	54,000	0.982	0.015	0.927	0.972	0.994	1.012
C9.perDF	54,000	0.981	0.016	0.922	0.971	0.994	1.015
C10.ncp	54,000	0.018	0.015	−0.012	0.006	0.028	0.073
C10.perDF	54,000	0.019	0.016	−0.015	0.006	0.029	0.078
delta.csq	54,000	82.830	66.497	6.626	35.338	114.514	358.427
rmsea.path	54,000	0.032	0.020	0.000	0.017	0.046	0.087
myMc	54,000	0.985	0.017	0.910	0.976	0.997	1.024
myTLI	54,000	0.987	0.018	0.867	0.979	0.997	1.072
target.csq	54,000	300.752	130.918	142.494	218.407	326.362	911.655
rmsea.afi	54,000	0.014	0.010	0	0.01	0.02	0

Mc, and χ^2, the interaction between measurement quality and severity of misspecification produced η^2 0.4 and above; while, an η^2 of 0.18 was observed for RMSEA. The rest of the higher-order effects were less than 0.06 for all global measures of fit, with χ^2 being the exception. In the context of a covariance misspecification, these findings held with a few notable exceptions. Namely, the impact of measurement quality was larger than the severity of the misspecification for all global measures of fit, with the exception being TLI. Specifically, η^2 was observed to be 0.03 for measurement quality, compared with η^2 0.6 and above for the rest; and η^2 was noticeably smaller for severity (0.27) for TLI compared with the other global measures of fit ($\eta^2 > 0.5$). With respect to the structural measures of fit regardless of the type of misspecification,

the resulting η^2 estimates were identical for the C9 and C10 SFIs controlling for whether they were estimated using the non-centrality parameter or ratio approach. Further, the η^2 estimates for the C9 and C10 SFIs were nearly identical. The main effect of misspecification severity was larger than any other main effect with $\eta^2 > 0.9$ for mean misspecifications and η^2 ranging from 0.77 to 0.84 for covariance misspecifications, with the exceptions being $\triangle\chi^2$ and RMSEA-P, which were found to be impacted by measurement quality more than severity (η^2 of 0.86 versus 0.84, $\triangle\chi^2$; η^2 of 0.82 versus 0.80, RMSEA-P). Overall, measurement quality was found to have a larger impact given a covariance misspecification (η^2 of 0.49 to 0.87), compared with a mean misspecification ($\eta^2$0.22 to 0.71). Interestingly, reference group sample size had a larger impact on structural measures (η^2 of 0.06 to 0.12), regardless of the type of misspecification, relative to the global measures of fit. On the other hand, no meaningful impact of model size was found across the types of misspecifications. In terms of higher-order effects, the measurement quality by severity interaction was larger for the covariance misspecification conditions (η^2 of 0.27 to 0.71) compared with mean misspecification conditions (η^2 of 0.15 to 0.60). With respect to the reference group sample size by severity interaction, this effect was found to be meaningful in the presence of a mean misspecification (η^2 of 0.6), whereas for a covariance misspecification η^2 was below 0.06 for all structural measures of fit. Another difference observed across the types of misspecifications was a meaningful 3-way interaction between measurement quality, reference group sample size, and severity which was observed for $\triangle\chi^2$ when it was a covariance misspecification, η^2 was 0.06.

Hit rates. After determining hit rates using the cut-offs from the Type I Error simulation, summary statistics were computed to assess the overall hit rate across all conditions for both the global and structural measures of fit; they are presented below. When evaluating the overall performance of the global measures of fit, it was observed that hit rates were higher when the mean structure was misspecified, compared with when the covariance structure was misspecified. For instance, considering an α of 0.05, TLI was able to detect mean misfit 47 percent of the time compared to 25 percent of

Table 17

Hit Rates: Global Measures

Note. μ *= hit rate;* σ *= standard deviation of hit rate; 95 =* α *of 0.05; 91 =* α *of 0.01; tli = Tucker-Lewis index; mc = McDonald's measure of centrality; chisq =* χ^2*; *.afi = global measure of fit; N = total number of replications*

Type of Misspecification:		Mean Structure		Covariance Structure	
Statistic	N	μ	σ	μ	σ
tli.afi.95	54,000	0.466	0.499	0.253	0.435
mc.afi.95	54,000	0.587	0.492	0.422	0.494
rmsea.afi.95	54,000	0.705	0.456	0.566	0.496
chisq.95	54,000	0.677	0.468	0.535	0.499
tli.afi.99	54,000	0.279	0.448	0.065	0.246
mc.afi.99	54,000	0.477	0.499	0.312	0.463
rmsea.afi.99	54,000	0.608	0.488	0.441	0.497
chisq.99	54,000	0.594	0.491	0.423	0.494

the time when the covariance structure was misspecified. A pattern observed that was expected was that power rates were lower when using an α of 0.01. For instance, the power rate for TLI went from 0.47 to 0.28 when it was a mean misspecification; similarly, power went from 0.25 to 0.07 when it was a covariance misspecification. For the most part, the variability in the expected power rates was similar, regardless of the choice for α, with the exception of TLI when it was a covariance misspecification. Specifically, when α was 0.05, the variability around the TLI hit rate was 0.49, compared with 0.25 when α was 0.01. Overall, the RMSEA possessed the most power to detect a mean misspecification at 0.71 and a power of 0.57 to detect a covariance misspecification. For the summary statistics of hit rates for the global measures of fit across both mean and covariance misspecifications, see Table 17.

Table 18

Hit Rates: Structural Measures

Note. μ *= hit rate;* σ *= standard deviation of hit rate; 95 =* α *of 0.05; 91 =* α *of 0.01; ncp = non-centrality parameter; pd =* $\frac{T_{ML}}{df}$*; delta.chisq =* $\triangle\chi^2$*; rmsea.p = RMSEA-P; N = total number of replications*

Type of Misspecification:		Mean Structure		Covariance Structure	
Statistic	N	μ	σ	μ	σ
c9.pd.95	54,000	0.778	0.416	0.754	0.431
c9.ncp.95	54,000	0.769	0.421	0.743	0.437
c10.pd.95	54,000	0.778	0.416	0.754	0.431
c10.ncp.95	54,000	0.769	0.421	0.743	0.437
rmsea.p.95	54,000	0.817	0.387	0.760	0.427
delta.chisq.95	54,000	0.827	0.378	0.771	0.420
c9.pd.99	54,000	0.696	0.460	0.597	0.491
c9.ncp.99	54,000	0.690	0.463	0.583	0.493
c10.pd.99	54,000	0.696	0.460	0.597	0.491
c10.ncp.99	54,000	0.690	0.463	0.583	0.493
rmsea.p.99	54,000	0.743	0.437	0.661	0.473
delta.chisq.99	54,000	0.753	0.431	0.677	0.468

Overall, the structural measures of fit were found to outperform the global measures of fit. Specifically, the range in hit rates for the structural measures of fit, with α of 0.05, ranged from 0.77 for the C9 and C10 SFIs to 0.817 for the RMSEA-P. As expected, these hit rates decreased when using an α of 0.01, with the C9 and C10 power rates decreasing to 0.69 and RMSEA-P decreasing to 0.74. Similar to the global measures of fit, the performance of the structural measures was greater when attempting to detect a mean misspecification. Across the board, the structural measures of fit out performed the global measures of fit, regardless of the type of misspecification. For instance, with an α of 0.05, power rates did not drop below 0.77 when the mean structure was misspecified or 0.74 when the covariance structure was misspecified. Overall, the $\triangle\chi^2$ had 0.83 power to detect a mean misspecification and 0.77 power to detect a covariance misspecification; these were the highest observed. Table 18 contains the sufficient statistics of the hit rates for all of the structural measures of fit across both mean and covariance misspecifications.

Impact of design factors on hit rates. Using the identical factorial ANOVA design as before, the impact of the study design factors on performance was examined. Regardless of the cut-off used (i.e., α of 0.05 or 0.01), the same pattern emerged, with the only difference being larger η^2 estimates for those defined using α of 0.01.

In the context of a mean misspecification, power rates were impacted by the main effect of measurement quality and severity, with η^2 estimates being larger for the former. Interestingly, when the misspecification was in the covariance structure, the η^2 estimates were greater for measurement quality, compared with when the mean structure was misspecified. On the other hand, η^2 estimates were smaller for the main effect of severity when the covariance structure was misspecified, compared to when the mean structure was misspecified. The sole exception was the TLI, where the inverse was true (i.e., larger η^2 estimates, given a mean misspecification). See Tables 19 and 20 for these η^2 estimates across all measures of fit when the mean and covariance structures, respectively, were misspecified.

Table 19

Effect of Design Factors on Power: Mean Misspecification

Note. $\alpha = 0.05$*; partial* $\eta^2 < 0.01$ *are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); ref.n = reference group sample size; tli = Tucker-Lewis index; mc = McDonald's measure of centrality; chisq =* χ^2*; *.afi = global measure of fit; ncp = non-centrality parameter; delta.csq =* $\triangle\chi^2$*; rmsea.p = RMSEA-P*

	tli.afi.95	mc.afi.95	rmsea.afi.95	chisq.95	c9.ncp.95	c10.ncp.95	rmsea.p.95	delta.chisq.95
pF (A)								
MQ (B)	0.340	0.134	0.054	0.061	0.019	0.019	0.039	0.038
ref.n (C)	0.016							
severity (D)	0.523	0.712	0.575	0.617	0.611	0.611	0.467	0.439
B:D	0.138	0.199	0.033	0.058	0.039	0.039	0.069	0.069
B:C:D	0.029							

Table 20

Effect of Design Factors on Power: Covariance Misspecification

Note. $\alpha = 0.05$*; partial* $\eta^2 < 0.01$ *are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); ref.n = reference group sample size; tli = Tucker-Lewis index; mc = McDonald's measure of centrality; chisq =* χ^2*; *.afi = global measure of fit; ncp = non-centrality parameter; delta.csq =* $\triangle\chi^2$*; rmsea.p = RMSEA-P*

	tli.afi.95	mc.afi.95	rmsea.afi.95	chisq.95	c9.ncp.95	c10.ncp.95	rmsea.p.95	delta.chisq.95
pF (A)								
MQ (B)	0.146	0.450	0.334	0.346	0.058	0.058	0.284	0.268
ref.n (C)								
severity (D)	0.240	0.430	0.325	0.338	0.453	0.453	0.324	0.315
B:D	0.060	0.263	0.085	0.101	0.039	0.039	0.147	0.144
C:D	0.014							
B:C:D			0.010					

Depending on the measure of fit, η^2 estimates differed for the main effects and the interaction effect. Therefore, a series of plots was generated to graphically represent the effect study design factors had on power rates and to what extent power rates differed in the presence of a misspecified mean or covariance structure. Each plot corresponds to hit rates determined using an α of 0.05 and are organized in an identical fashion. The x-axis contains the severity of the misspecification, with power along the y-axis.

McDonald's measure of centrality. When evaluating global model fit using the Mc index, its ability to detect either a misspecified mean or covariance structure was found to be impacted by measurement quality. In terms of the sample size for the reference group, this factor was not found to be meaningful. As the severity of the misspecification increases power rates increased. This pattern was observed with one exception; specifically, when the covariance structure was misspecified and measurement quality was poor, power rates remained similar across severity levels and did not surpass 0.25. With a misspecified covariance structure, Mc Reached the nominal power rate of 0.8 when attempting to detect a large misspecification. On the other hand, when Mc attempts to detect a moderately misspecified mean structure, nominal levels are reached with moderate levels of measurement quality. Mc power rates depend on the type of misspecification and the level of measurement quality. Ultimately, Mc performs best when measurement quality is high and the mean structure is misspecified (Figure 10).

Tucker-Lewis index. The performance of the TLI was unexpected. It was observed that as measurement quality decreases, power rates increase. This was true whether the mean or covariance structure was misspecified. The only instance observed in which TLI had acceptable power was when measurement quality was poor and the mean structure was moderately misspecified, or when the covariance structure was largely misspecified, given balanced group sample sizes. Interestingly, when measurement quality was high, TLI was powered to detect a large mean misspecification given balanced groups (Figure 11).

Root mean squared error of approximation. The RMSEA is impacted by measurement quality more in the context of a misspecified covariance structure, than a mean

misspecification. For example, when measurement quality is poor, the RMSEA is not powered to detect a small, medium, or large covariance misspecification (power is 0.5 or lower), while, the RMSEA is powered, regardless of measurement quality to detect a large misspecification. When the mean structure is moderately misspecified, RMSEA had adequate power given moderate or high levels of measurement quality. On the other hand, when the covariance structure is moderately misspecified, the RMSEA is powered when groups are balanced and standardized loadings are 0.6 (Figure 12).

χ^2. The global fit statistic was found to perform better when detecting a moderate misspecification in the mean structure compared with the covariance structure. It also was apparent that measurement quality had a large impact on χ^2 when attempting to detect a covariance misspecification. For this statistic, reference group sample size did not have a large impact; this is hypothesized to be due to the overall sample size remaining the same at 2000 across all simulation condition (Figure 13).

Non-centrality parameter based C9. When evaluating a misspecified mean structure, the C9 SFI was found not to be impacted by measurement quality or reference group sample sizes. As expected, when the severity of the misspecification increased, so did power. Nominal power levels were observed, given a moderately misspecified mean structure. On the other hand, when attempting to detect a moderately misspecified covariance structure, standardized factor loadings needed to be 0.6 or higher. When measurement quality was poor, the C9 SFI was adequately powered to detect a largely misspecified covariance structure with a power around 0.9 (Figure 14).

Non-centrality parameter based C10. Unsurprisingly, the C10 SFI performed in an identical fashion (Figure 15).

Root mean squared error of approximation - Path. When the covariance structure was misspecified, the RMSEA-P was adequately powered to detect all possible levels of misspecification, given that measurement quality was high. When measurement quality was medium or poor, the RMSEA-P was adequately powered to detect a medium or large effect, respectively, in the covariance structure. When the mean structure was misspecified, RMSEA-P was powered to detect a medium or a large misspecification,

regardless of measurement quality levels. Therefore, it is clear that the RMSEA-P is sensitive to measurement quality when the misfit is in the covariance structure (Figure 16).

$\triangle\chi^2$. The performance of the traditional nested model test was observed to be similar to that of the RMSEA-P. That is, measurement quality had a larger impact when the covariance structure was misspecified and it was adequately powered to detect a small misspecification in the covariance structure when measurement quality was high (Figure 17).

Simultaneous Misspecifications. Table 21 contains the summary statistics for both the global and structural measures of fit. Overall, the measures of fit indicate a worse fitting model as compared with when only either the mean or covariance structure was misspecified. This was expected due to the conditions detailed here correspond to instances in which the structural model is grossly misspecified. For instance, when only the covariance or mean structure was misspecified, the mean C10-NCP SFI was estimated to be 0.982 and 0.969, respectively. compared with 0.948 when these structures were simultaneously misspecified.

Impact of design factors on measures of fit. With respect to the global measures of fit, the main effects of measurement quality and mean misspecification severity were substantial. For instance, with respect to Mc, the partial η^2 estimates for these factors were 0.797 (MQ) and 0.804 (severity-mean). Interestingly, partial η^2 estimates were larger for the measurement quality factor than the covariance misspecification severity factor. In terms of the main effect of model size, these partial η^2 estimates were found to be larger than those for the reference group sample size factor. With respect to higher-order interaction effects, large partial η^2 estimates were observed for the 2-way interactions between measurement quality and mean misspecification severity, and measurement quality and covariance misspecification severity. See Appendix A for all relevant partial η^2 estimates.

With respect to the structural measures of fit, the main effect of mean misspecification severity was found to have an obvious impact. Unlike the global measures of fit, the

Table 21

Descriptive Statistics: Simultaneous Misspecification

*Note. ncp = non-centrality parameter; perDF = $\frac{T_{ML}}{df}$; delta.csq = $\triangle\chi^2$; rmsea.path = RMSEA-P; myMc = McDonald's measure of centrality; myTLI utilize the manually specified baseline model; target.csq = χ^2; *.afi = global measure of fit.*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
C9.ncp	162,000	0.948	0.030	0.816	0.927	0.972	1.011
C10.ncp	162,000	0.052	0.030	−0.011	0.028	0.073	0.184
delta.csq	162,000	190.928	128.426	7.843	87.831	267.467	765.211
rmsea.path	162,000	0.057	0.024	0	0.04	0.1	0
myMc	162,000	0.958	0.031	0.819	0.939	0.982	1.024
myTLI	162,000	0.966	0.033	0.738	0.956	0.988	1.077
target.csq	162,000	569.998	253.238	151.746	322.689	761.998	1,388.866
rmsea.afi	162,000	0.021	0.010	0.000	0.014	0.027	0.058

measurement quality design factor was found not to be statistically significant for all structural measures, with RMSEA-P and $\triangle\chi^2$ being the exceptions; for these measures, measurement quality was the most significant design factor. All structural measures were impacted to some degree by covariance misspecification severity, reference group sample size, and model size. With respect to the higher-order interactions, the following were statistically significant: mean and covariance misspecification severity (RMSEA-P), measurement quality and mean severity (all SFMs), and measurement quality and reference group sample size ($\triangle\chi^2$).

For all relevant partial η^2 estimates, see Appendix A.

Hit rates. As previously mentioned, these conditions contained the most severe misspecifications and, therefore, it was expected for the measures of fit to possess notable statistical power to detect the misspecifications. Table 22 contains the overall

hit rates for both the global and structural measures of fit using an α of 0.05 and 0.01. As expected, the overall hit rates were observed to be smaller when consulting cut-offs using an α of 0.01, compared with 0.05. With an α level of 0.05, all measures of fit had a power of at least 0.80 to detect the structural misfit, with one exception, TLI. The observed power for TLI was found to be 0.547, which was well below the remaining measures of fit. When using an α of 0.01, the global measures of fit, except for TLI (0.286), were just under 0.80, whereas, power rates for all structural measures of fit were 0.916 (C9 and C10) or above.

Table 22

Hit Rates: Simultaneous Misspecification

Note. μ *= hit rate;* σ *= standard deviation of hit rate; 95 =* α *of 0.05; 99 =* α *of 0.01; tli = Tucker-Lewis index; mc = McDonald's measure of centrality; chisq =* χ^2*; *.afi = global measure of fit; N = total number of replications*

Type I Error Rate:		$\alpha = 0.05$			$\alpha = 0.01$	
Measure	N	μ	σ	Measure	μ	σ
tli.afi.95	162,000	0.547	0.498	tli.afi.99	0.286	0.452
mc.afi.95	162,000	0.817	0.387	mc.afi.99	0.729	0.445
rmsea.afi.95	162,000	0.818	0.386	rmsea.afi.99	0.706	0.456
chisq.95	162,000	0.850	0.357	chisq.99	0.776	0.417
c9.ncp.95	162,000	0.970	0.171	c9.ncp.99	0.916	0.277
c10.ncp.95	162,000	0.970	0.171	c10.ncp.99	0.916	0.277
rmsea.p.95	162,000	0.967	0.178	rmsea.p.99	0.943	0.232
delta.chisq.95	162,000	0.970	0.172	delta.chisq.99	0.947	0.224

Impact of design factors on hit rates. Tables 23 and 24 contain the partial η^2 estimates for the global and structural measures of fit, respectively. These tables illustrate the extent to which study design factors influenced the performance of the measures and

their ability to correctly reject the model. As before, it appears that the main effect of mean severity is larger than that of covariance severity across the two types of measures (global and structural). For the global measures of fit, partial η^2 estimates ranged from 0.285 to 0.431 for mean severity, and, these estimates ranged from 0.10 to 0.17 for covariance severity. On the other hand, partial η^2 estimates for the structural measures of fit ranged from 0.08 to 0.10 for mean severity, and these estimates ranged from 0.04 to 0.05 for covariance severity. The effect of measurement quality was found to have a statistically significant effect on the global measures of fit, all corresponding to a large effect, while medium effects were found for only the RMSEA-P and the $\triangle\chi^2$, with estimates of 0.07 and 0.06, respectively. With respect to model size, no statistically significant effects were observed for the structural measures of fit, whereas a medium effect was observed for the TLI with a partial η^2 of 0.20. In terms of reference group sample size, this design factor was found to be negligible for both the global and the structural measures of fit.

In terms of higher order interactions, a pattern emerged across the global and structural measures of fit. Specifically, the interaction between measurement quality and mean severity was statistically significant for all measures except the C9 and C10 SFIs, with partial η^2 ranging from 0.06 (TLI) to 0.13 (Mc and RMSEA). The interaction between mean severity and covariance severity also was found to be statistically significant for all measures of fit, with TLI being the exception; partial η^2 ranged from 0.08 ($\triangle\chi^2$) to 0.12 (Mc). With respect to the TLI, the interaction between model size and measurement quality was statistically significant (partial $\eta^2 = 0.7$). In fact, a statistically significant four-way interaction was observed for TLI concerning model size, measurement quality, mean severity, and covariance severity, resulting in a partial η^2 of 0.06; therefore, it was not surprising to observe a partial η^2 of 0.11 for the interaction between model size, measurement quality, and mean severity. An additional interaction effect was found to be statistically significant for the conventional structural measures of fit which was the three-way interaction between measurement quality, mean severity, and covariance severity, with the partial η^2 estimates of 0.09 and 0.08 for the RMSEA-P

Table 23

Effect of Design Factors on Power: Simultaneous Misspecification - Global

Note. $\alpha = 0.05$*; partial* $\eta^2 < 0.01$ *are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); ref.n = reference group sample size; sev.mean = severity of mean misspecification; sev.var = severity of covariance misspecification; tli = Tucker-Lewis index; mc = McDonald's measure of centrality; chisq =* χ^2*; *.afi = global measure of fit*

	tli.afi.95	mc.afi.95	rmsea.afi.95	chisq.95
pF (A)	0.198	0.032	0.049	
MQ (B)	0.401	0.211	0.199	0.161
ref.n (C)	0.013			
sev.mean (D)	0.431	0.337	0.325	0.285
sev.var (E)	0.173	0.118	0.121	0.100
A:B	0.073	0.019	0.011	
A:D	0.017	0.020	0.024	
B:D	0.059	0.134	0.127	0.125
A:E			0.012	
B:E	0.016	0.011	0.012	0.011
D:E	0.044	0.121	0.112	0.111
A:B:D	0.110	0.010	0.011	
A:B:E	0.025			
A:D:E	0.015			
B:D:E	0.033	0.048	0.039	0.031
A:B:D:E	0.064	0.010	0.017	
A:B:C:D:E	0.013			

and $\triangle\chi^2$, respectively.

Table 24

Effect of Design Factors on Power: Simultaneous Misspecification - Structural

Note. $\alpha = 0.05$*; partial* $\eta^2 < 0.01$ *are left blank and empty columns are removed. MQ = measurement quality; pF = model size (p:f); ref.n = reference group sample size; sev.mean = severity of mean misspecification; sev.var = severity of covariance misspecification; ncp = non-centrality parameter; delta.csq =* $\triangle\chi^2$*; rmsea.p = RMSEA-P*

	c9.ncp.95	c10.ncp.95	rmsea.p.95	delta.chisq.95
pF (A)				
MQ (B)	0.014	0.014	0.066	0.061
ref.n (C)				
sev.mean (D)	0.078	0.078	0.101	0.092
sev.var (E)	0.053	0.053	0.047	0.044
A:D	0.010	0.010	0.016	0.015
B:D	0.027	0.027	0.121	0.113
B:E	0.011	0.011	0.046	0.045
D:E	0.101	0.101	0.089	0.084
A:B:D			0.015	0.015
B:D:E	0.022	0.022	0.086	0.084

To better understand the impact of the design factors on the measures of fit, a series of figures (one per measure of fit) are presented below. Each figure contains 3 plots, with power on the y-axis and mean severity on the x-axis. The top plot corresponds to models with a small misspecification to the covariance structure. The bottom plot corresponds to models with a largely misspecified structure. As before, line color corresponds to measurement quality and line type corresponds to the reference group

sample size.

McDonald's measure of centrality. The performance of Mc in the context of simultaneous misspecifications to the mean and covariance structures was found to be affected by measurement quality. For instance, when measurement quality was high, Mc was adequately powered to reject the model for all conditions, except when the misspecification to the covariance structure was small. On the other hand, when measurement quality was poor, only when the mean severity was large was Mc adequately powered across all sample sizes (Figure 18).

Tucker-Lewis index. As before, the performance of the TLI was peculiar. Unlike the other measures of fit, as measurement quality decreased, statistical power increased. When measurement quality was high, the TLI was never powered to correctly reject the model. On the other hand, when measurement quality was low, the TLI approached 0.80 power across all levels of covariance misspecifications, given that the mean structure was moderately misspecified. When the covariance structure was severely misspecified, along with a mean structure that was misspecified to either a moderate or large degree, the TLI was powered to reject the model when measurement quality levels were moderate (Figure 19).

Root mean squared error of approximation. When the mean and covariance structures were simultaneously misspecified, the RMSEA was found to perform quite well. When measurement quality was high or moderate, nominal power levels were reached, given a small misspecification to the covariance structure and a medium misspecification to the mean structure. When measurement quality was poor, a large misspecification to the mean structure was required across all levels of covariance misspecification in order for the RMSEA to be adequately powered. Given medium misspecification to the covariance structure, a small, medium, and large misspecification to the mean structure was required for the RMSEA to be powered to reject the model when measurement quality was high, moderate, and poor, respectively. When the covariance structure was misspecified to a large degree, the RMSEA was powered to reject the model across all levels of mean severity when measurement quality was high or moderate (Figure 20).

χ^2. The test statistic was found to perform in an identical manner to the RMSEA; however, its power levels were observed to be higher than the RMSEA (Figure 21).

Noncentrality based C9 and C10. Both the C9 and C10 structural fit indices performed in a similar manner. Given a covariance structure that was misspecified to a medium degree, regardless of mean severity, measurement quality, or reference group sample size, the C9 and C10 SFIs were powered to correctly reject the model. When the covariance structure was misspecified to a small degree, nominal power levels were reached when the mean severity was small (high measurement quality) or medium (moderate and poor measurement quality). Altogether, power levels never dropped below 0.6 (Figures 22 and 23).

Root mean squared error of approximation - Path. The RMSEA-P was found to be statistically powered to reject models with either a medium or large misspecification of the covariance structure, regardless of measurement quality or reference group sample size. On the other hand, when the covariance structure was misspecified to a small degree, nominal power levels were reached when measurement quality was either moderate or high. Power levels did not drop below 0.4 across all simulation conditions. See Figure 24.

$\triangle\chi^2$. The conventional nested model test statistic was found to perform similarly as the RMSEA-P (Figure 25).

Global versus Structural Measures. In an effort to compare the performance of the different types of fit measures, an analysis of covariance was performed. The dependent variable was the hit rate at the 95^{th} percentile for the RMSEA, RMSEA-P, TLI, C9, χ^2, and $\triangle\chi^2$. It was decided to omit the C10 SFI due to its performance being identical to the C9 SFI. With respect to the Mc, it was decided that although it is an absolute measure of global fit, its very different from the others. This was done for each structural misspecification condition: mean, covariance, and simultaneous. Table 25 contains the partial η^2 estimates for the main effects and their interactions.

Mean misspecification. After controlling for the discrepancy fit value ($\hat{F}_{ML}$) partial η^2 estimates were 0.04 (Structural Indicator) or lower, and all corresponded to small effect

Table 25

Effect of Design Factors Controlling for $\hat{F}_{ML}$

Note. Partial $\eta^2 < 0.01$ *are left blank and empty columns are removed.*

$\hat{F}_{ML}$ *= minimum discrepancy value; Structural = indicates type of evaluation; Taxonomy = indicates type of inference (e.g., incremental); pF = model size (p:f); MQ = measurement quality*

	Mean	Covariance	Simultaneous
$\hat{F}_{ML}$	0.150	0.023	0.098
Structural (A)	0.043	0.098	0.106
Taxonomy (B)	0.021	0.024	0.037
pF (C)			0.098
MQ (D)		0.058	0.026
A:B		0.018	0.037
A:C			0.012
B:D	0.037	0.055	0.053
A:B:D	0.010	0.018	0.044

sizes, with the exception of $\hat{F}_{ML}$, which was 0.15. The next largest η^2 observed was for the interaction effect between type of measure (incremental, absolute, and test statistic) and measurement quality at 0.037.

Covariance misspecification. When the structural model was misspecified in the covariance structure, it was observed that $\hat{F}_{ML}$ had a small effect (0.023) compared to when the mean structure was misspecified (0.15). Larger effects were observed for all main and interaction effects when the covariance structure was misspecified with the largest corresponding to the effect of structural measures versus global measures: mean partial $\eta^2 = 0.043$ versus 0.098. Measurement quality also was observed to have a medium effect on hit rates when the covariance structure was misspecified ($\eta^2 = 0.058$), whereas, there was no statistically significant effect observed for the main effect of measurement quality on hit rates when the mean structure was misspecified. It follows then that a larger effect for the interaction between type of measure and measurement

quality was observed when the covariance structure was misspecified 0.037, compared with when the mean structure was misspecified.

Simultaneous misspecification. Not surprisingly, partial η^2 estimates were as large or larger for all main and interaction effects, compared with when only one structure was misspecified. The exceptions were for $\hat{F}_{ML}$ 0.098 versus 0.15 when the mean structure was misspecified, and the main effect of measurement quality 0.026 versus 0.058 when the covariance structure was misspecified. Interestingly, given simultaneous misspecifications, a statistically significant effect for model size was observed (partial $\eta^2 = 0.098$) and a small-to-medium effect for the three-way interaction between structural versus global, type of measure, and measurement quality (partial $\eta^2 = 0.053$) was observed.

Visualizing power rates. Figure 26 contains boxplots that correspond to power rates for each of the measures of fit. The boxplots are color-coded based on level of measurement quality (red = 0.4, green = 0.6, and blue = 0.8), with power on the y-axis and type of structural misspecification on the x-axis, and a dashed black line at power = 0.80. Controlling for the minimum discrepancy fit value, the conditional mean power rates favor structural measures of fit, regardless of whether the measure is a test statistic (χ^2), incremental (C9) or absolute (RMSEA-P) fit index. On the far right side of the figure, the boxplots for TLI are plotted; it is clear that this measure of fit performed poorly and displayed the opposite behavior from the other measures of fit; namely, as measurement quality decreased, power for this measure to correctly reject the misspecified model increased. When measurement quality is high and both the mean and covariance structures are misspecified, the range in statistical power for TLI is shocking: 0.0 to 0.50.

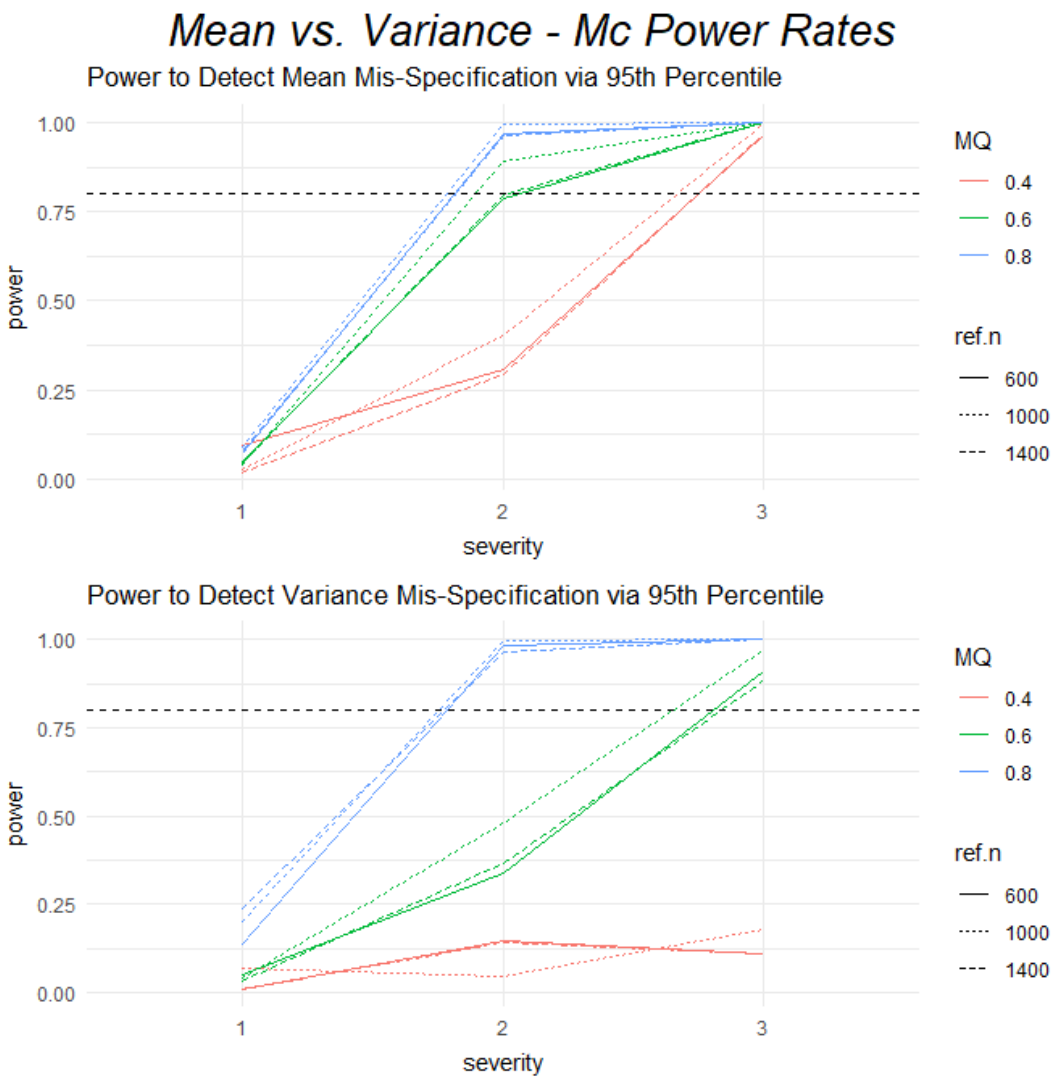

Figure 10. Power for Mc to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

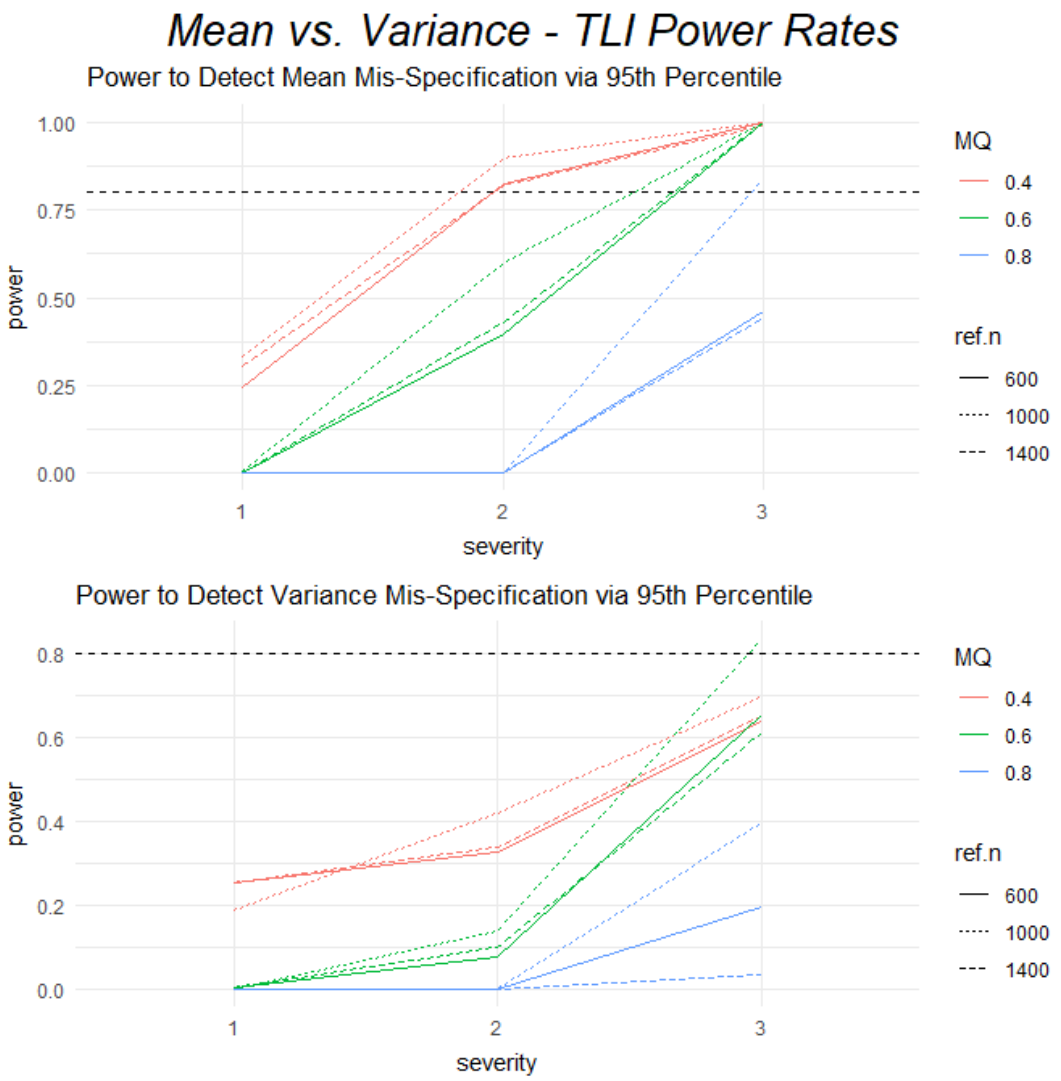

Figure 11. Power for TLI to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

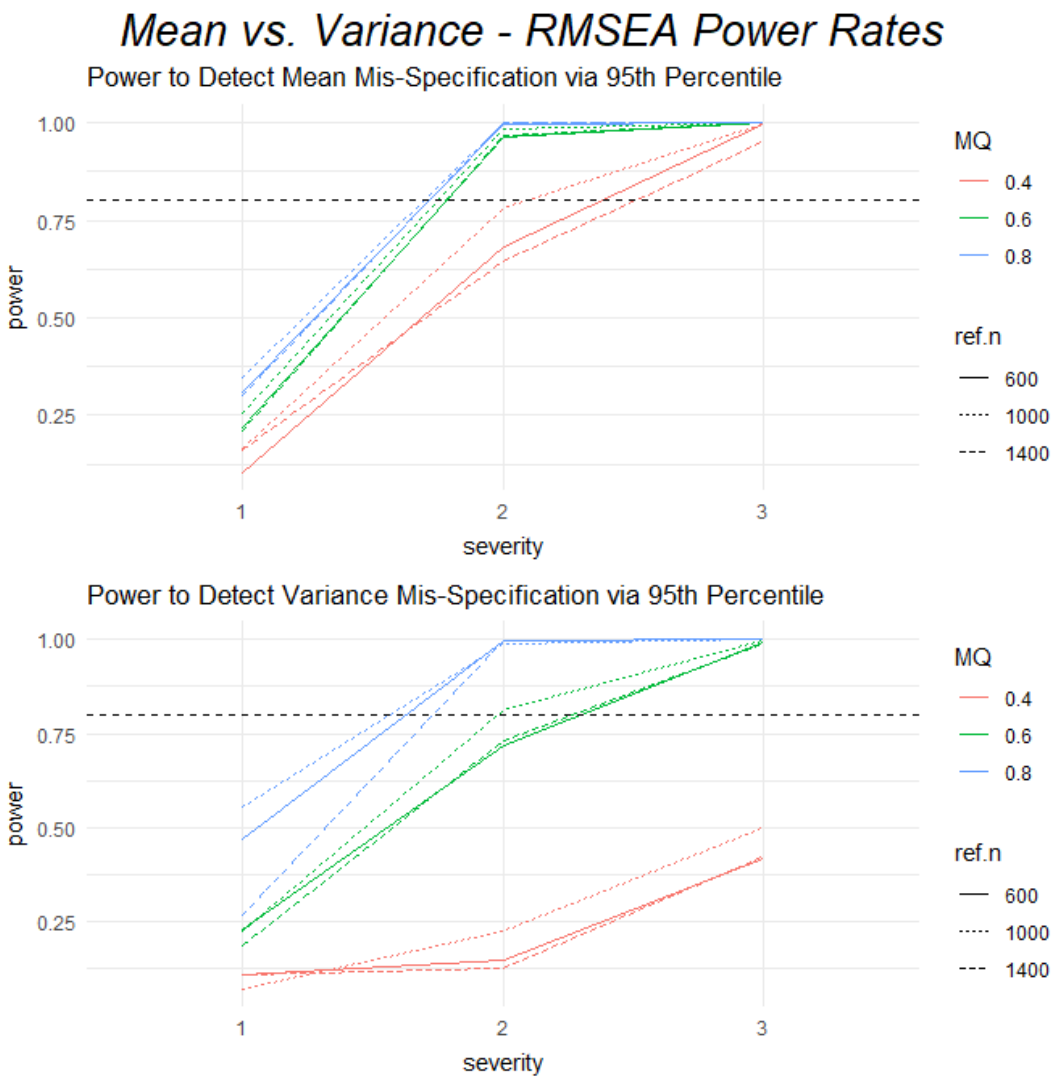

Figure 12. Power for RMSEA to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

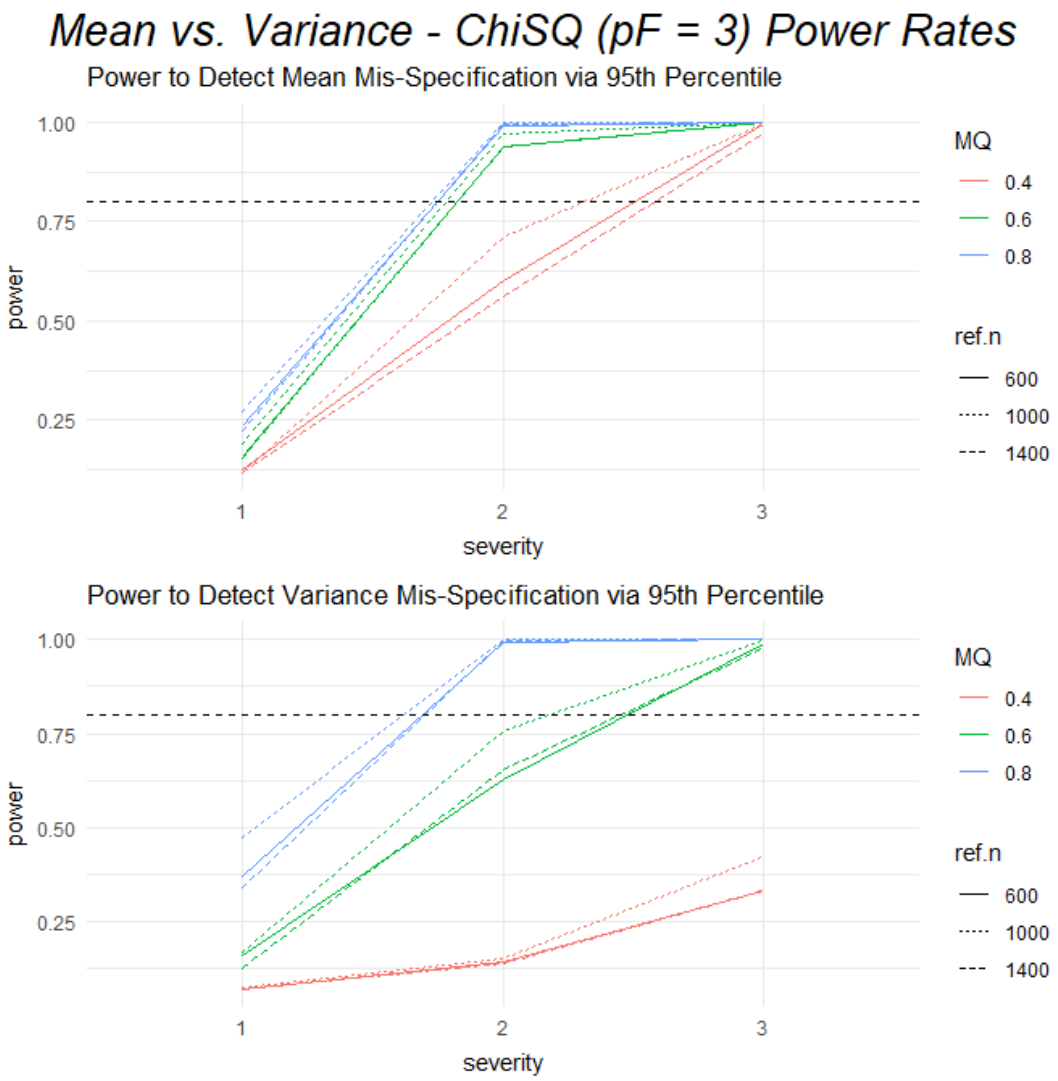

Figure 13. Power for χ^2 to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

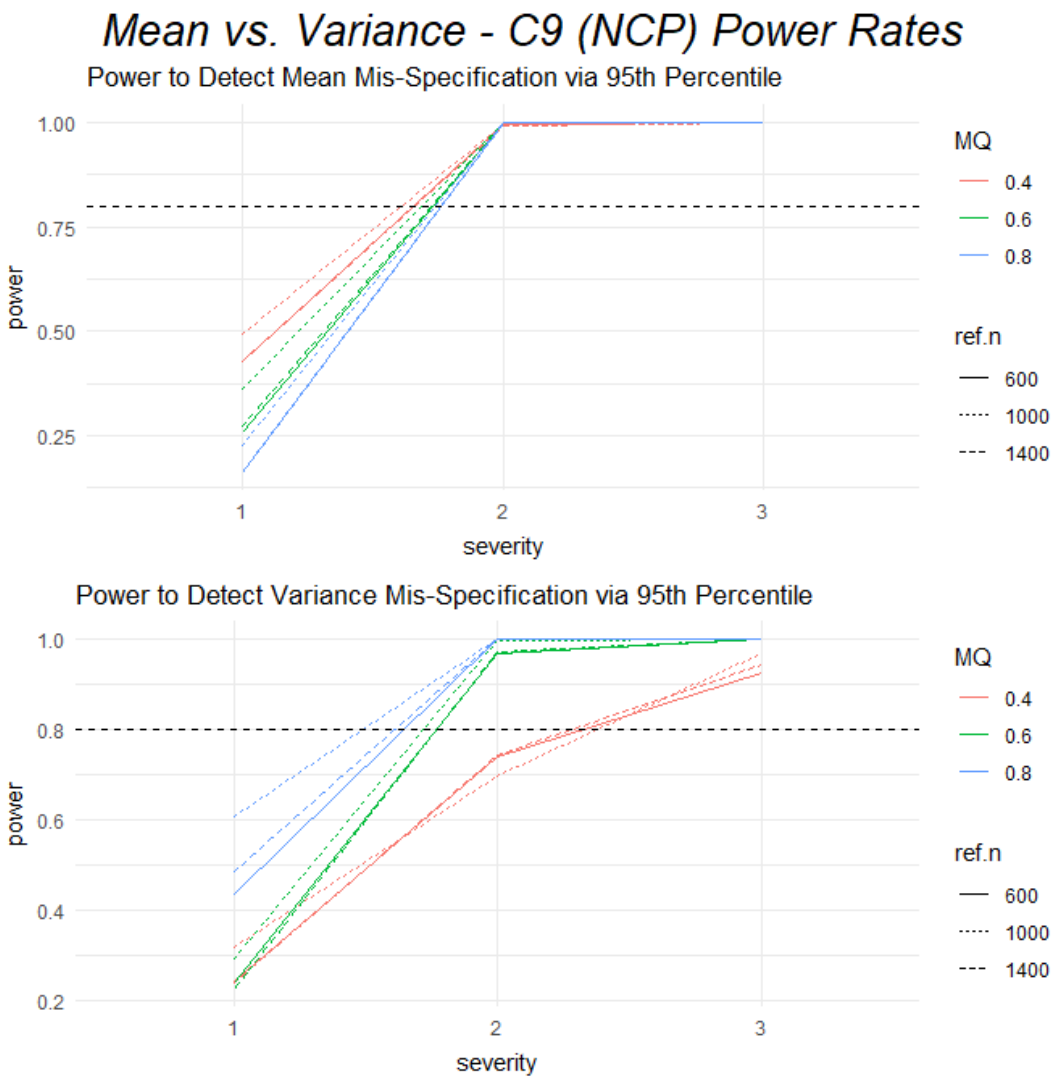

Figure 14. Power for C9 to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

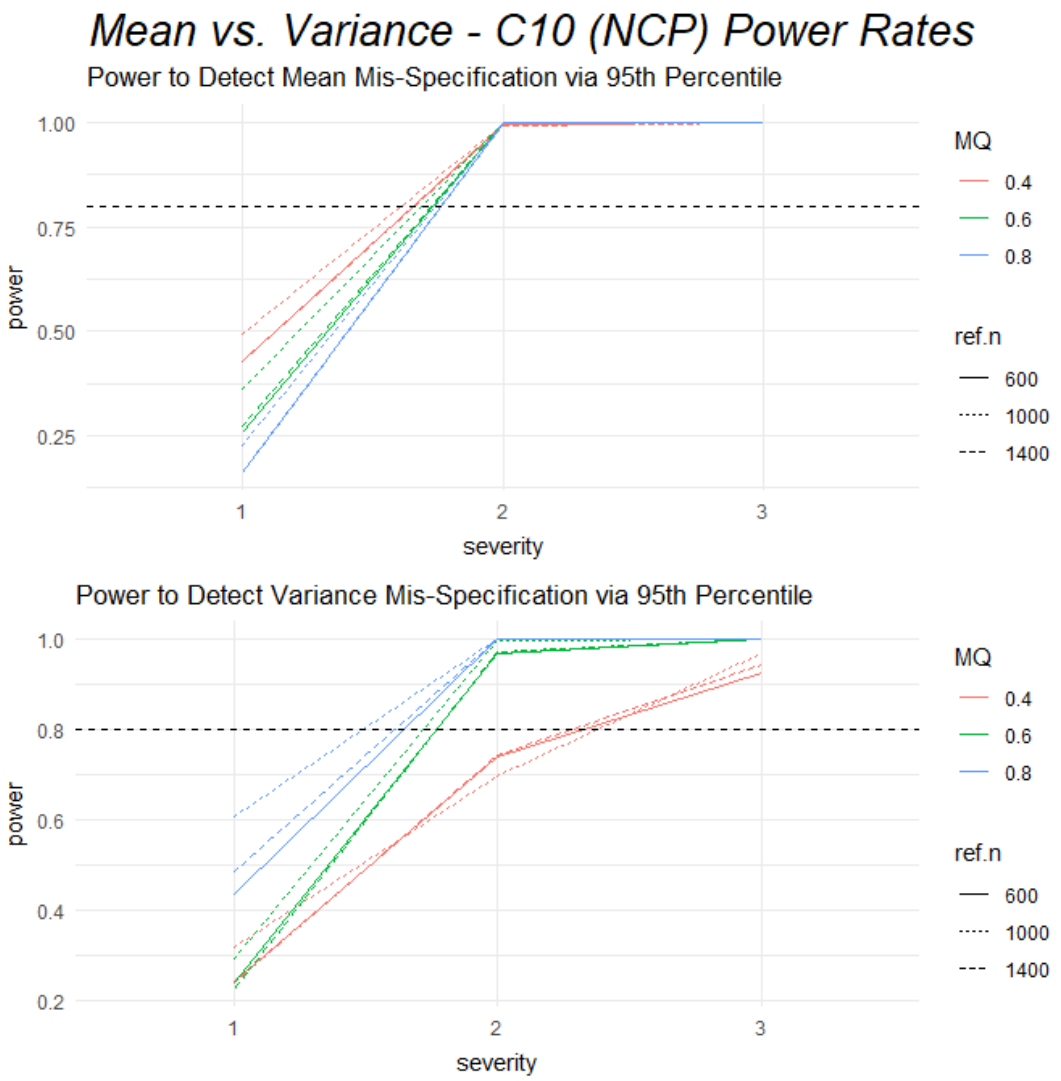

Figure 15. Power for C10 to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

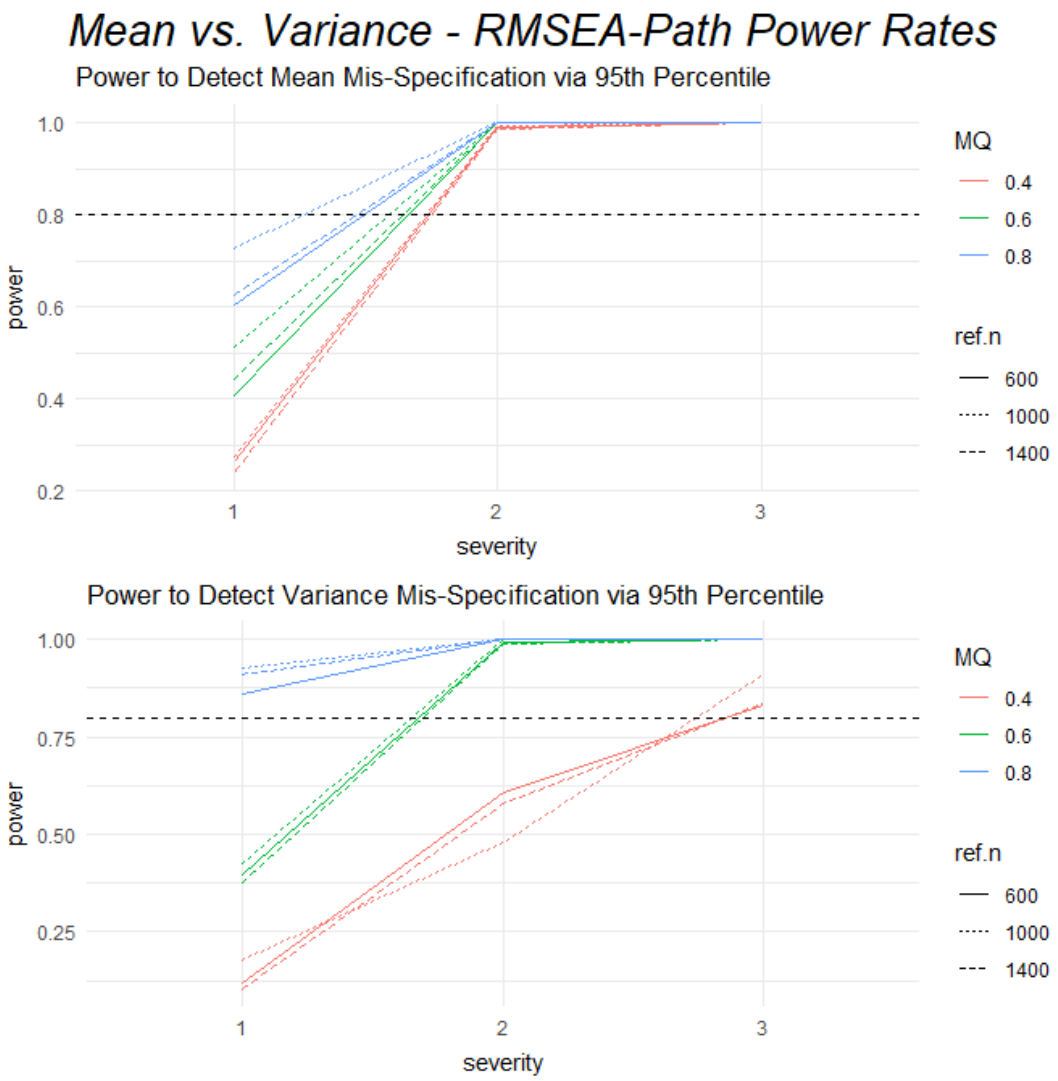

Figure 16. Power for RMSEA-P to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

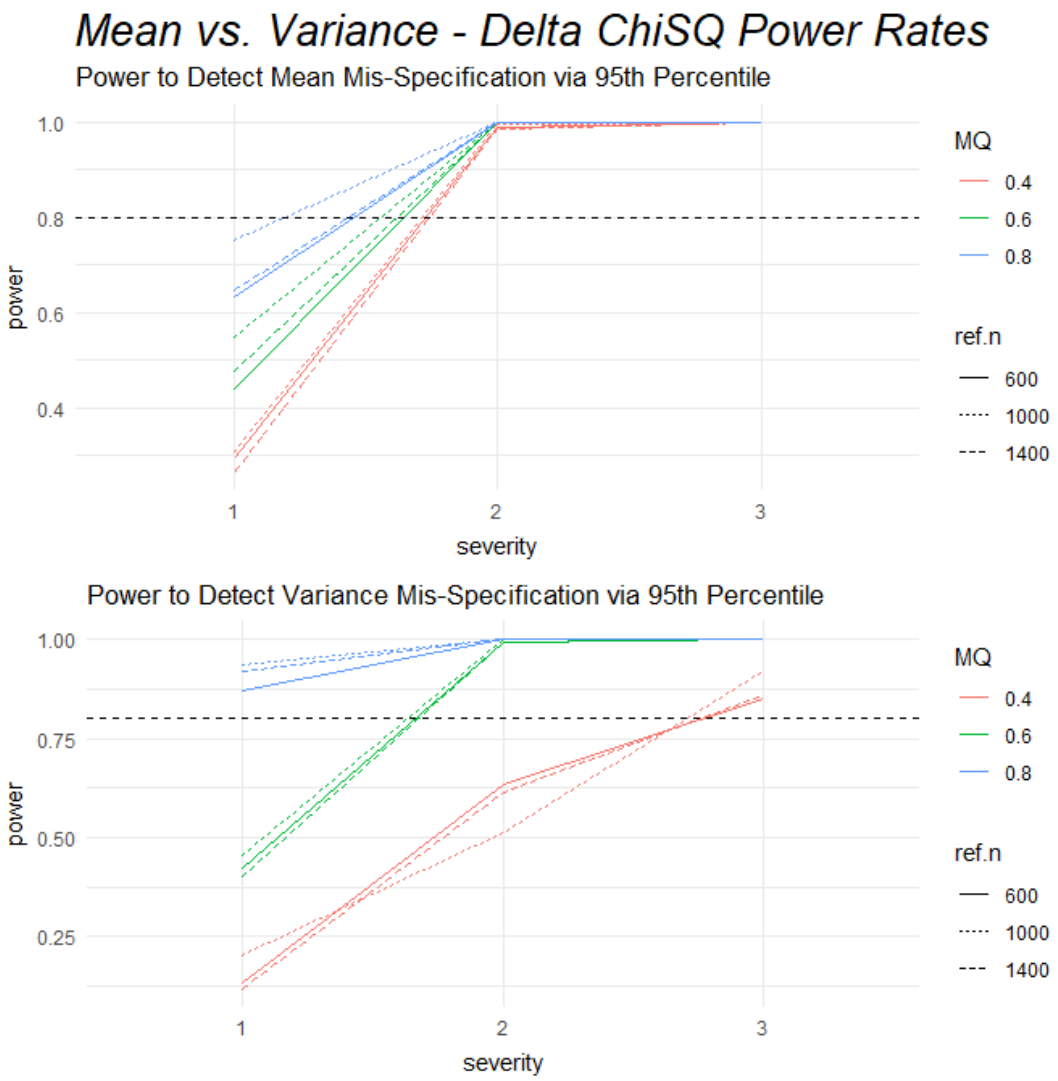

Figure 17. Power for $\triangle\chi^2$ to detect a mean or covariance misspecification assuming $\alpha = 0.05$ based on severity of misspecification, measurement quality, and reference group sample size

Note. MQ = measurement quality; ref.n = reference group sample size

Top corresponds to a misspecified mean structure and the plot on the bottom corresponds to when the covariance structure was misspecified.

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

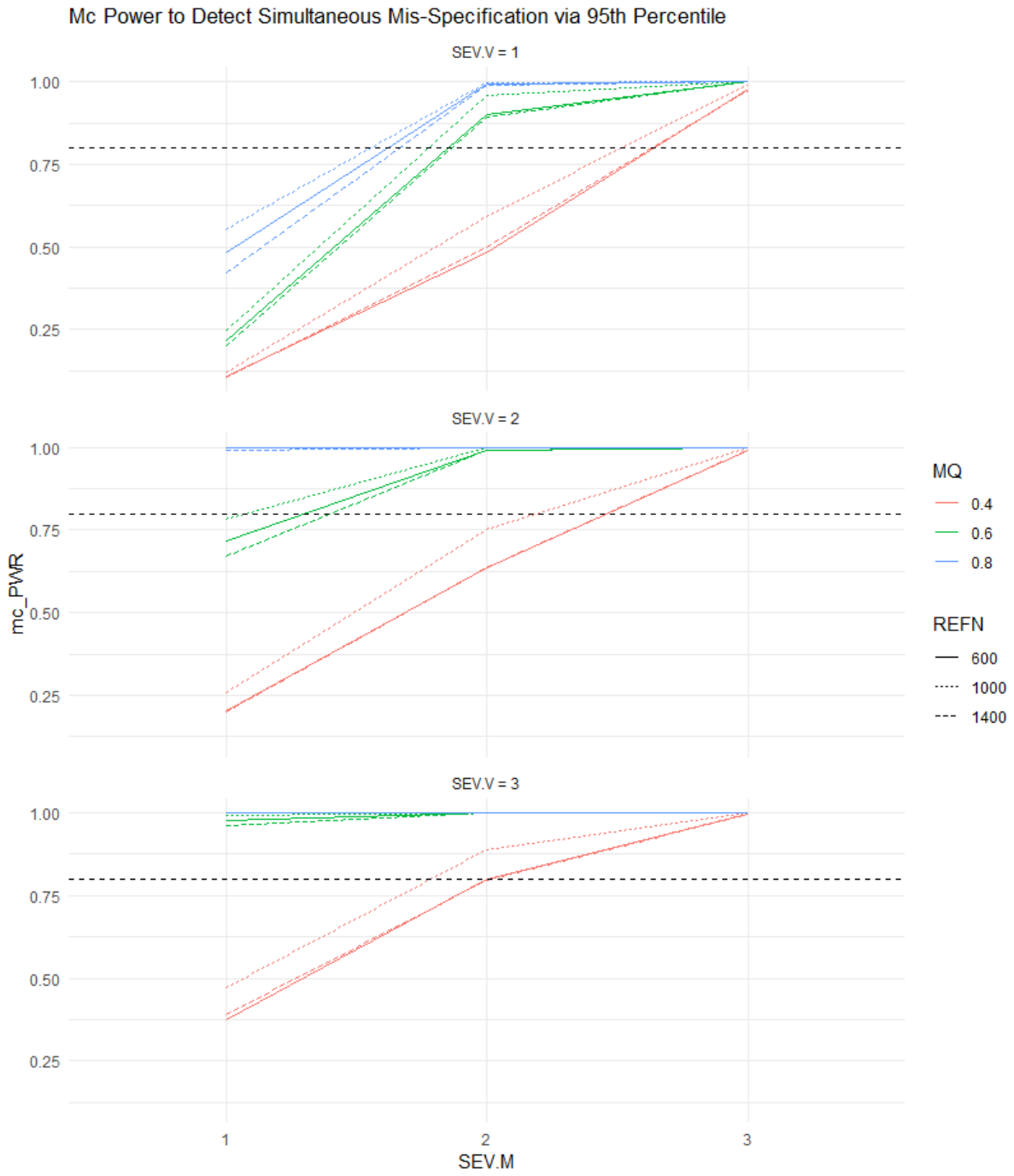

Figure 18. Power for Mc to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

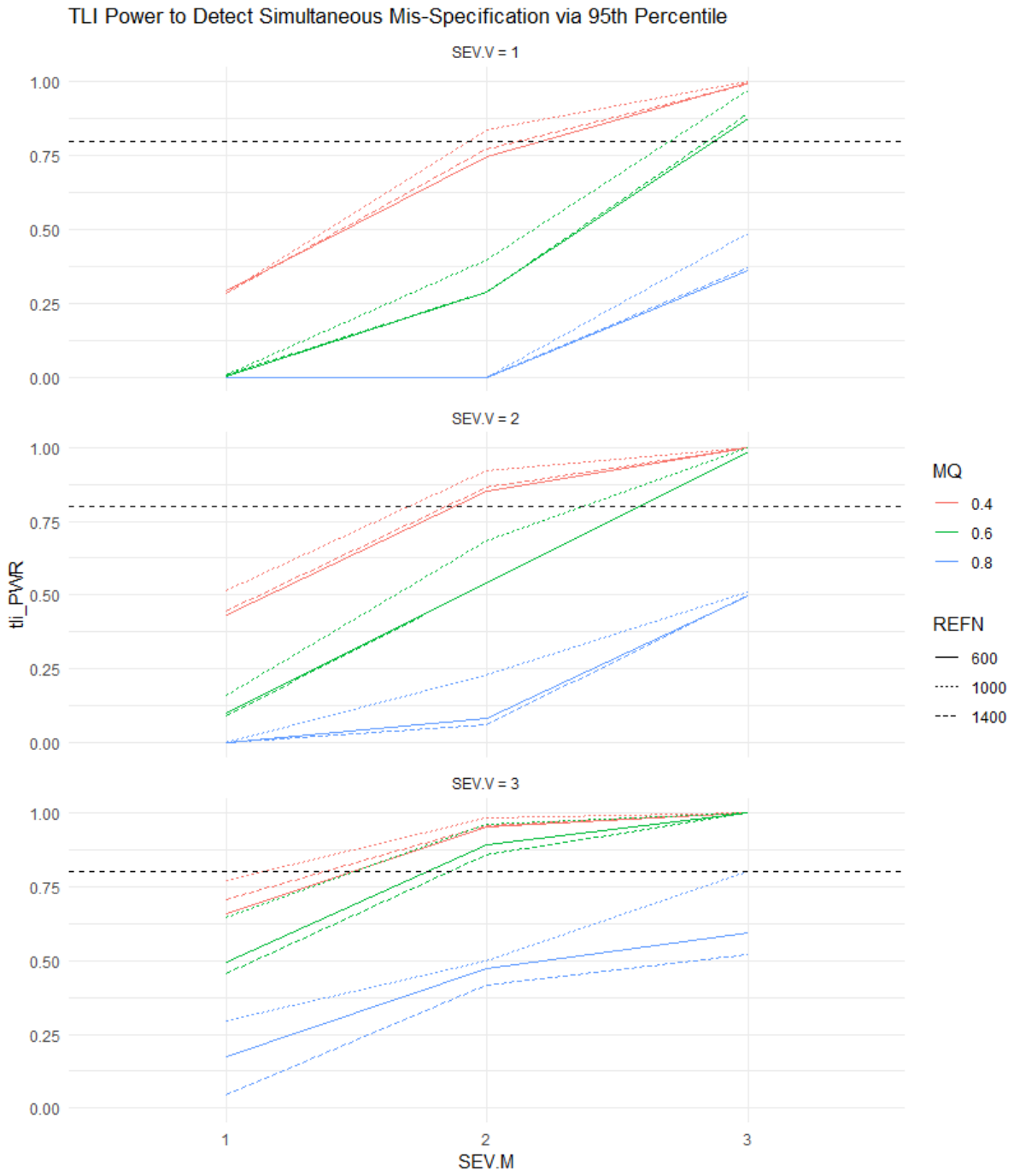

Figure 19. Power for TLI to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

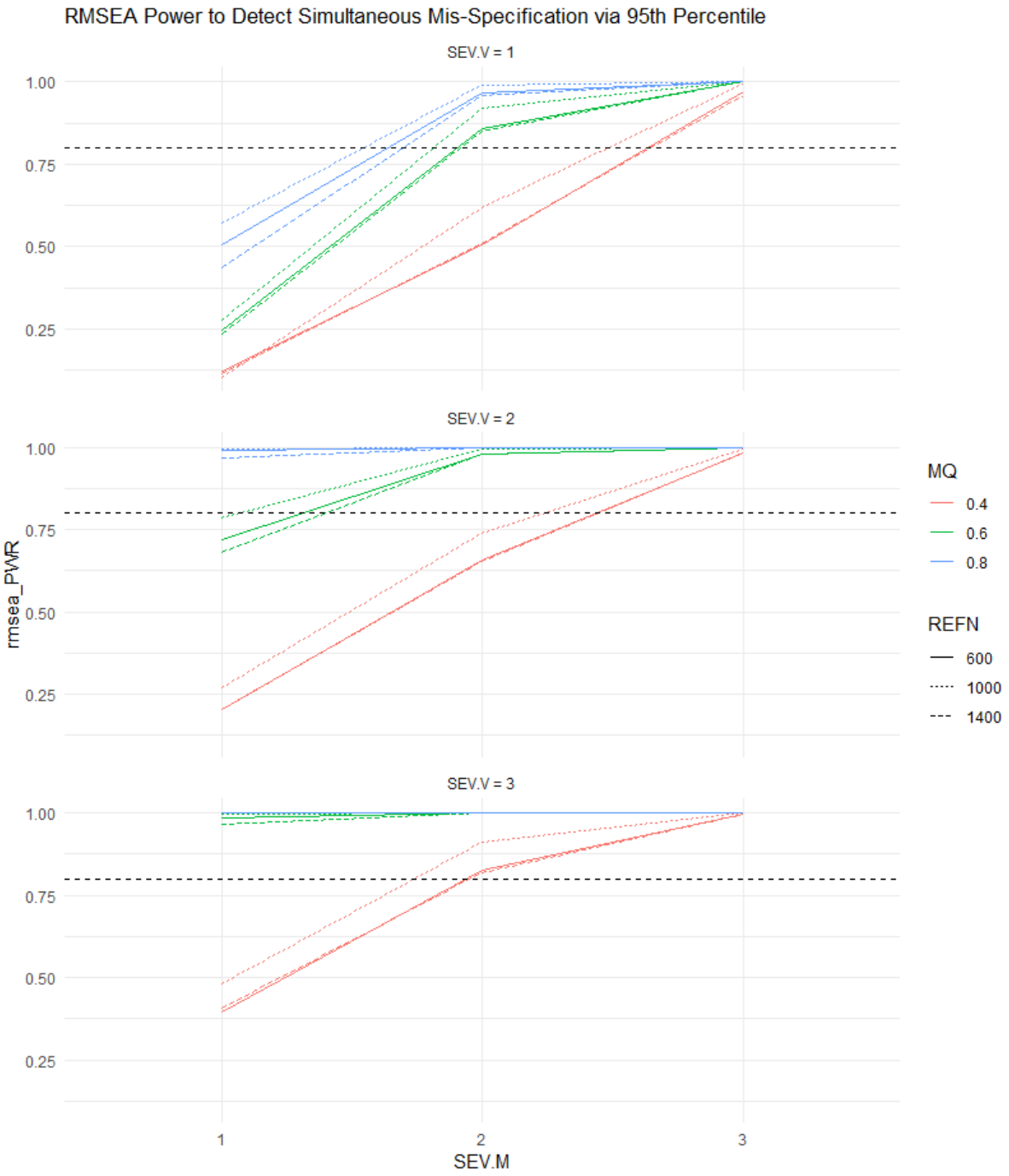

Figure 20. Power for RMSEA to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

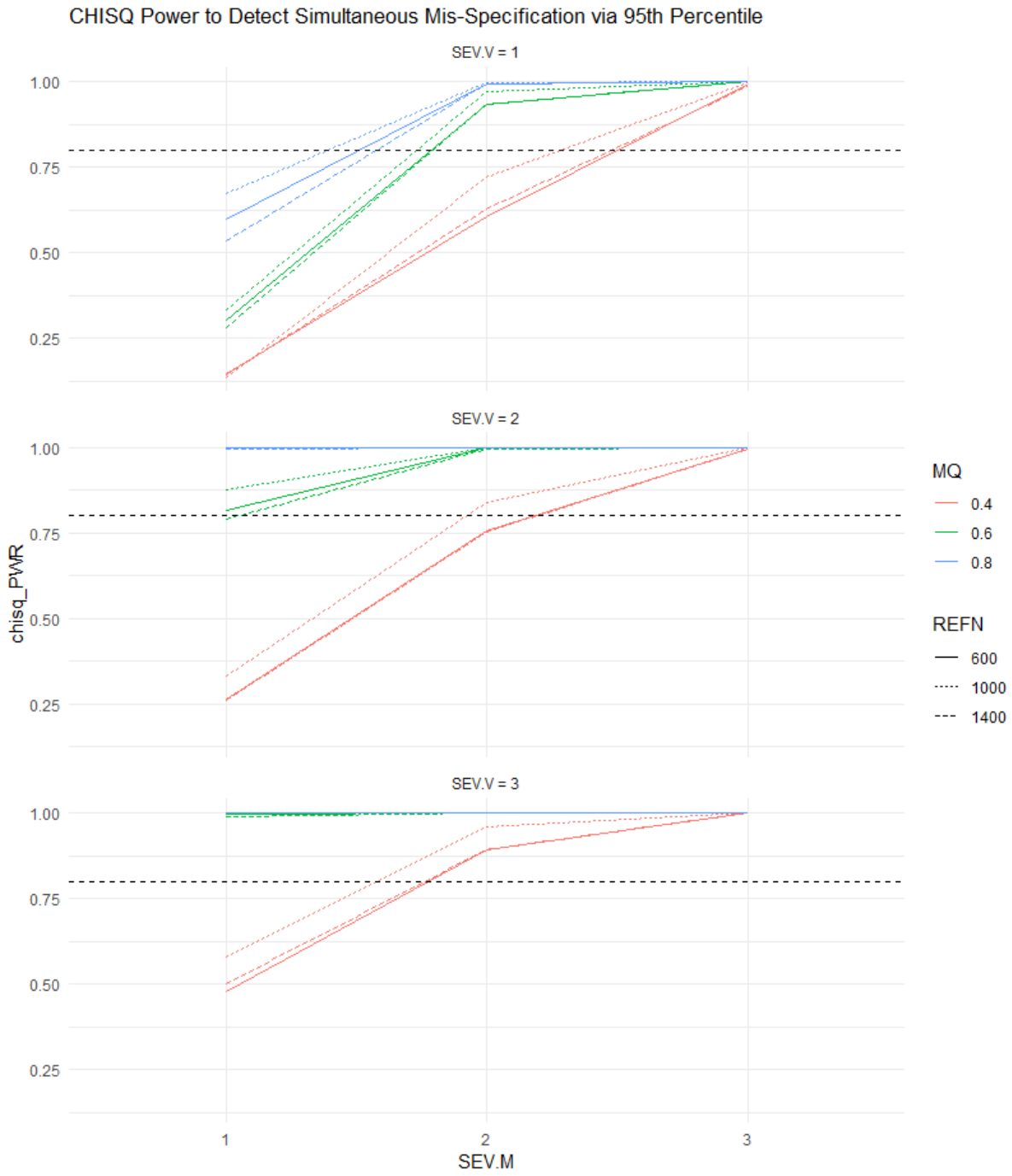

Figure 21. Power for χ^2 to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

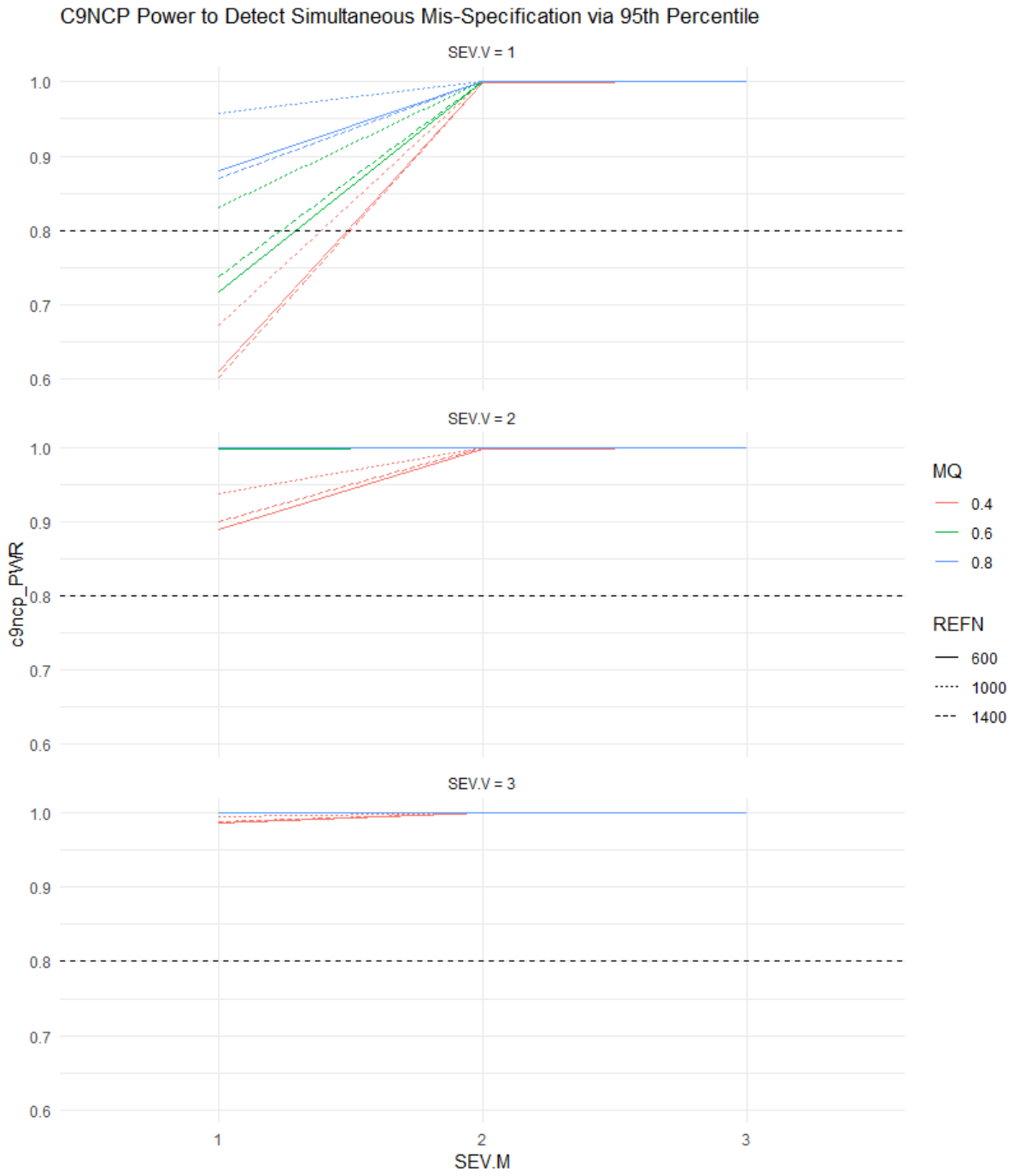

Figure 22. Power for C9 to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

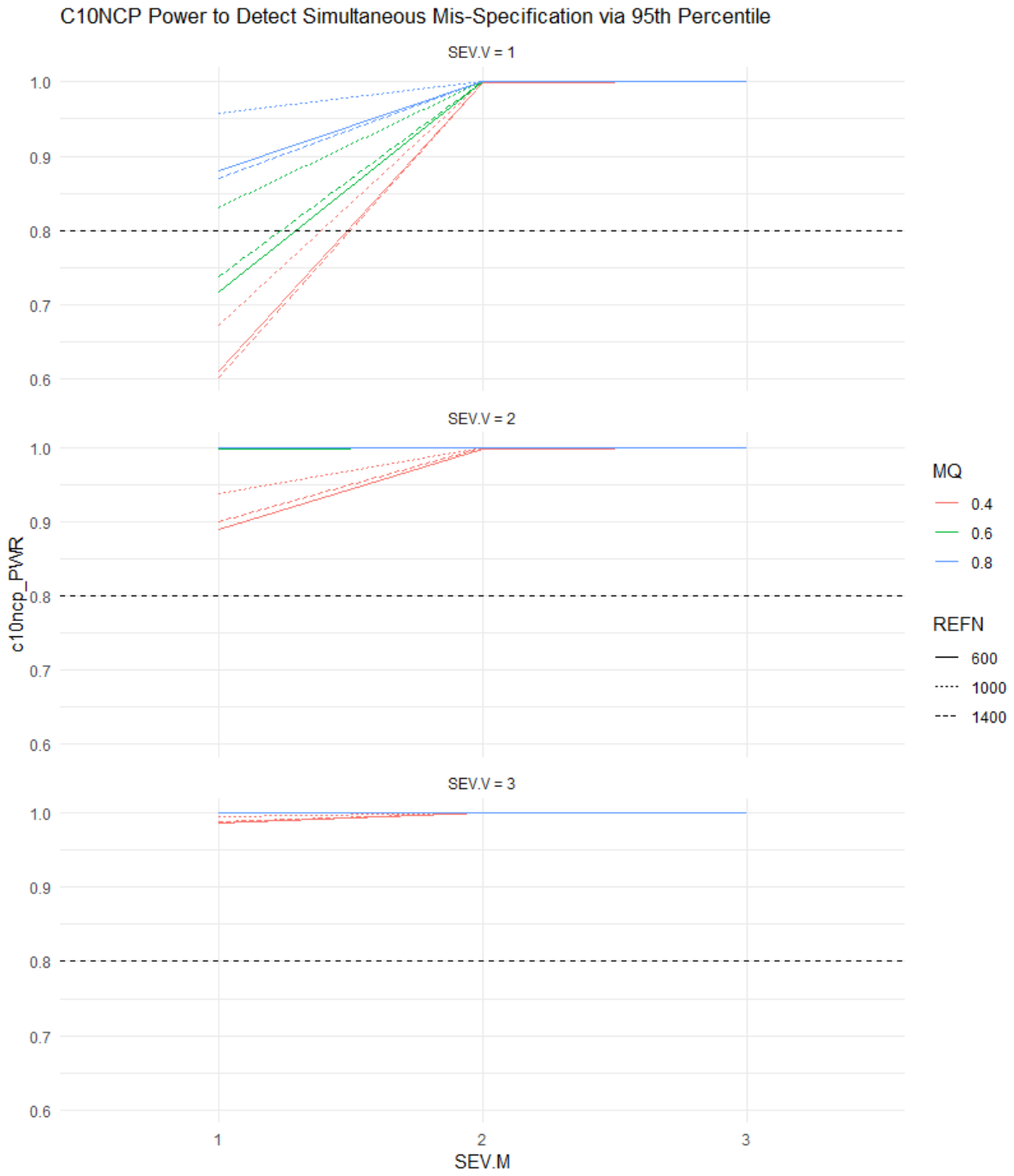

Figure 23. Power for C10 to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

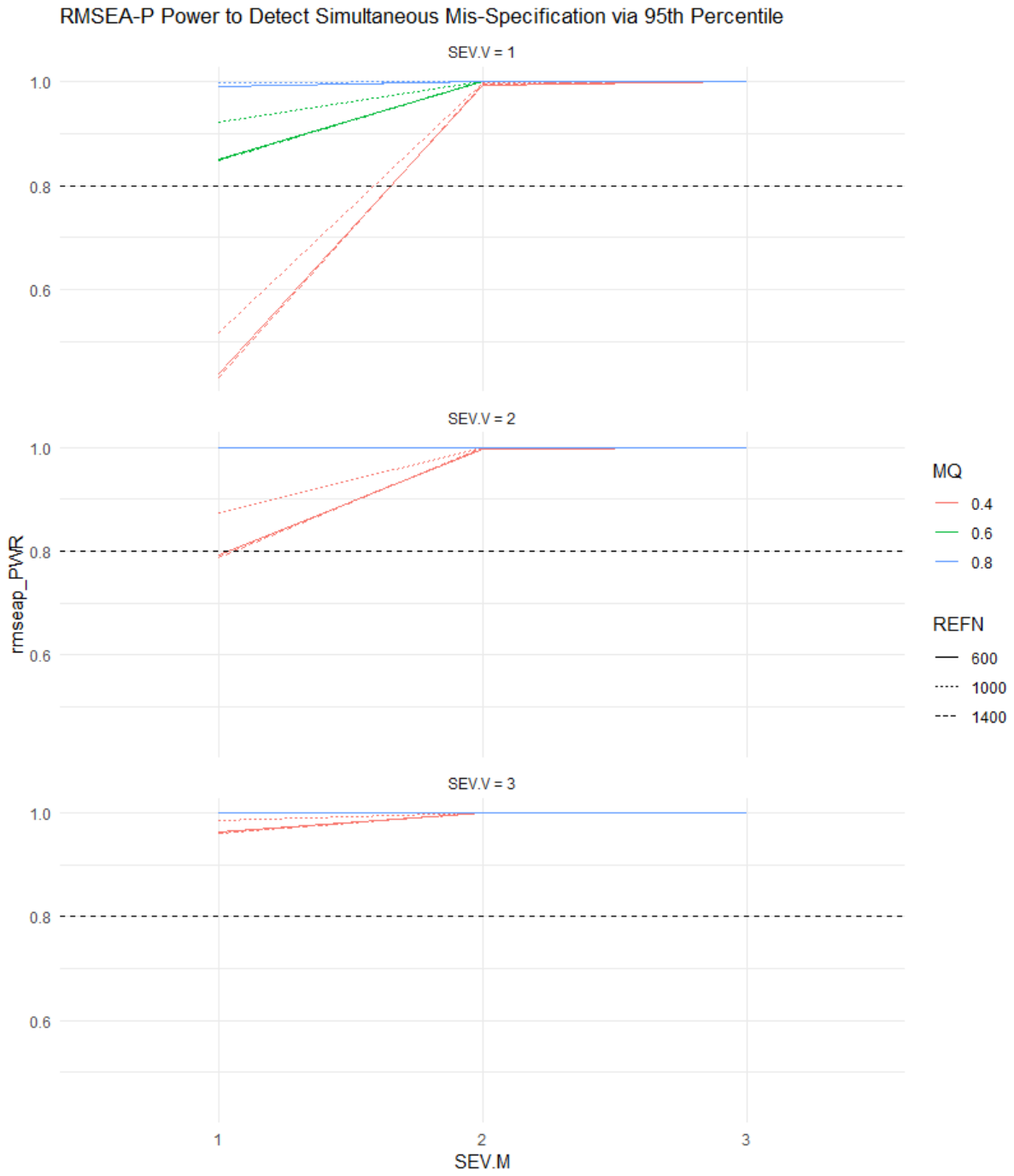

Figure 24. Power for RMSEA-P to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

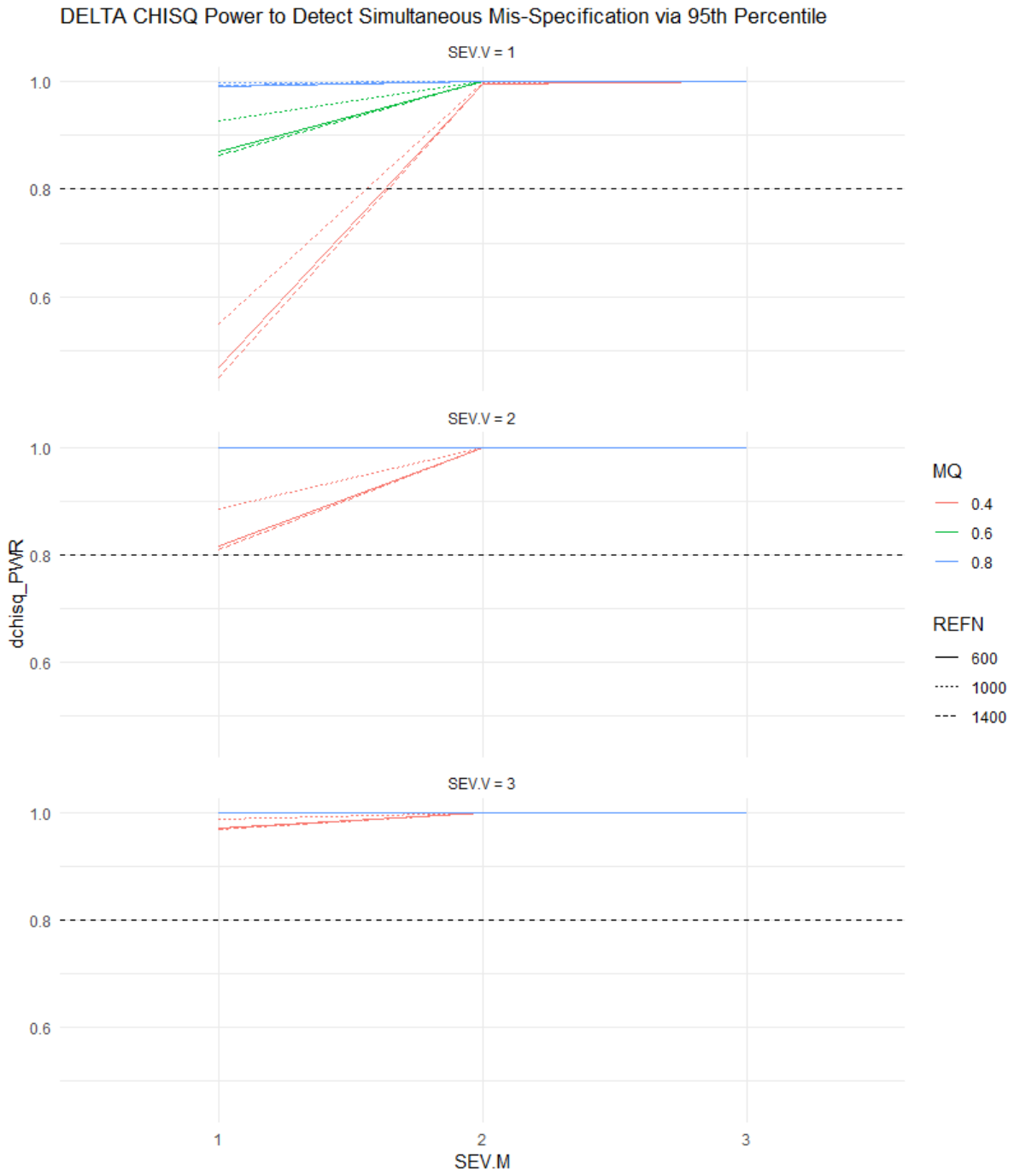

Figure 25. Power for $\triangle\chi^2$ to detect simultaneous misspecifications assuming $\alpha = 0.05$ based on measurement quality (MQ), and reference group sample size (ref.n)

Note. Plot corresponds to the severity of covariance misspecification

Line color: Red ($\Lambda = 0.4$), green ($\Lambda = 0.6$), and blue ($\Lambda = 0.8$).

Line Type: Solid line (ref.n = 600), dotted (ref.n = 1000), and dashed (ref.n = 1400)

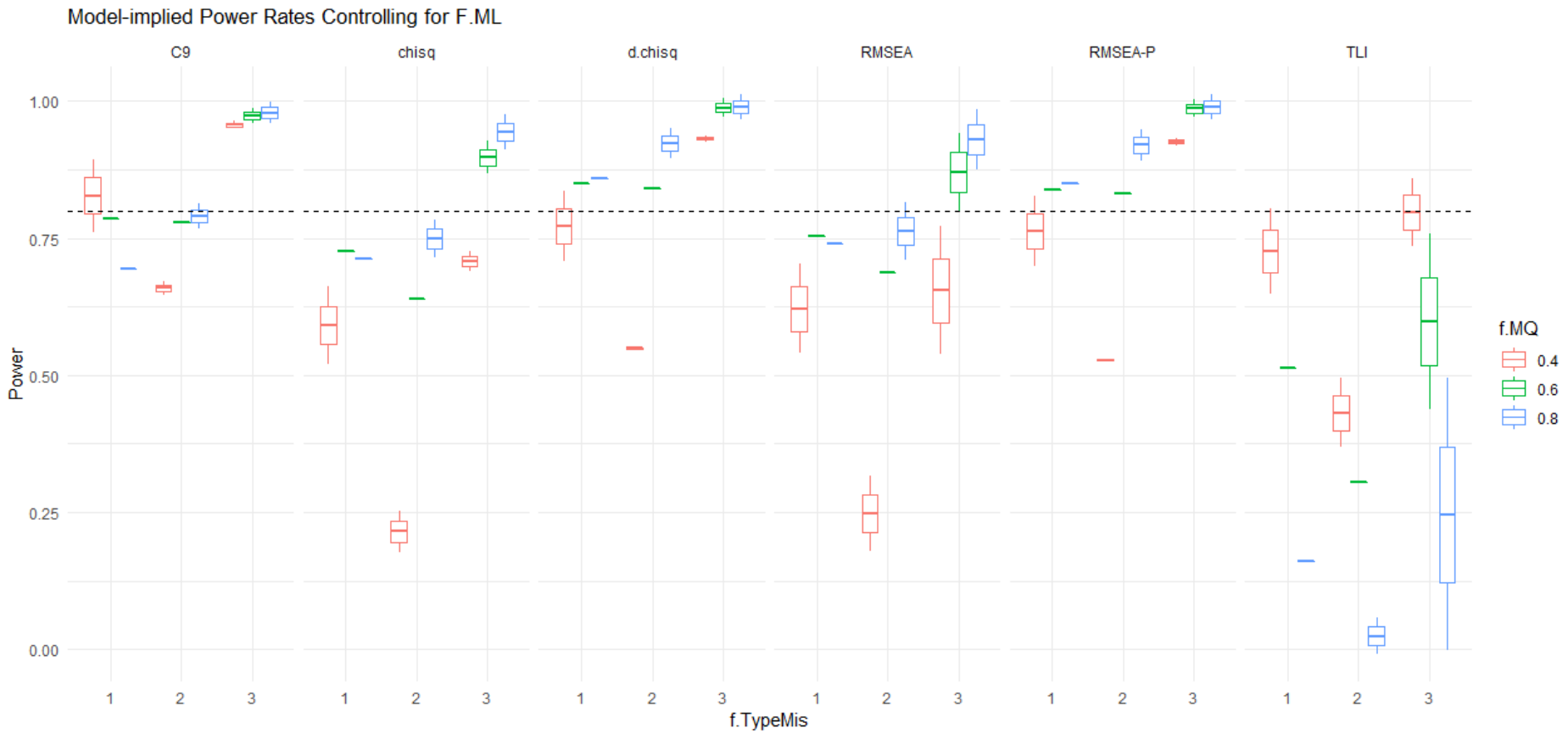

Figure 26. Power rates for measures controlling for $\hat{F}_{ML}$ X type of misspecification and measurement quality

Note. Boxplot: red ($\Lambda = 0.4$)*, green* ($\Lambda = 0.6$)*, and blue* ($\Lambda = 0.8$)

f.TypeMis: 1 (mean misspecification), 2 (covariance misspecification), 2 (simultaneous misspecification)

Summary of Results

Structural misspecifications were placed on either the mean or covariance structure independently in an effort to properly investigate the relative performance of the various global and structural measures of fit. In other words, it was possible to ascertain whether performance of the measures of fit was impacted by design factors and whether their influence depended on the type of structural misspecification. Overall, measures of fit were found to possess more statistical power and were more likely to correctly reject a model when the mean structure was misspecified than when the covariance structure was misspecified. Further, when the covariance structure was misspecified, measurement quality had a larger influence on the performance of the measures of fit, compared with when the mean structure was misspecified. As shown in Figures 10 through 17, the structural measures of fit outperformed the global measures of fit. Further, these plots showed the clear impact measurement quality had on power rates across both the structural and global measures of fit when a model's covariance structure was misspecified, compared with when the mean structure was misspecified. The performance of the TLI given either a misspecified mean or covariance structure was surprising because it was found to be unaffected by measurement quality in the Type I Error simulation. Further, measurement quality had an inverse effect on power rates for the TLI compared with all of the other measures of fit; specifically, as measurement quality decreased, power for the TLI increased. In fact, when measurement quality was high, power levels were all 0.80 or below, with the exception of a largely misspecified mean structure and balanced groups. TLI should not be utilized to assess structural model misfit.

As when only one structure was misspecified, given simultaneous misspecifications, structural measures were found to outperform the global measures of fit. Overall, the various measures of fit were found to possess more statistical power to correctly reject the misspecified model compared with when only the mean or covariance structure was misspecified. With respect to the effect design factors had on performance, the RMSEA and the Mc were observed to be impacted by the varying levels of measurement quality

and mean misspecification severity, whereas all measures of fit were impacted by varying levels of severity of mean and covariance misspecifications. Interestingly, the performance of the $\triangle\chi^2$ and RMSEA-P were found to be influenced by varying levels of measurement quality and severity of both mean and covariance misspecifications. Controlling for the discrepancy fit value, it was observed that measurement quality still had an impact on hit rates when the covariance structure was misspecified, and this effect carried over to conditions in which both the mean and covariance structures were misspecified. Moreover, depending on whether the measure was a fit statistic or an absolute/incremental fit index, measurement quality influenced power rated and this was true given either a covariance misspecification or simultaneous misspecifications. Overall, it was observed that the structural measures of fit outperformed the global measures of fit.

6. Discussion

Recently, approaches for evaluating structural model fit have been proposed. These approaches result in the estimation of structural fit indices (SFIs) in an effort to test the appropriateness of the structural model. As with the global fit indices (GFIs), it is unclear what constitutes a *good* fitting structural model, as the sampling distribution for the SFIs are unknown. The literature has shown that the cut-offs for the common GFIs provided by Hu and Bentler (1999) are susceptible to measurement quality; ultimately, these cut-offs determine the fate of a candidate model. Miles and Shevlin (2007) go so far as to say, "If you wish your model to fit, according to the proposed criteria, ensure your measures are unreliable" (p. 874).

Another aspect of reliability that creates disagreement between common GFIs is the number of indicators per latent variable (Ding et al., 1995). Specifically, as model size (or p:f) increases, so does construct reliability, as well as the burden on estimation. Specifically, the estimator is tasked with reproducing a larger variance-covariance matrix. As a result, the RMSEA tends to perform better given larger p:f ratios (Kenny et al., 2015); whereas, CFI and TLI perform worse (Ding et al., 1995).

Varying two facets of construct reliability, measurement quality and model size, this dissertation was conducted to examine the sampling distribution of the proposed SFIs when the estimation model was correct in the population (Type I Error simulation). Subsequently, the performance of the proposed SFIs was compared to the conventional $\triangle\chi^2$ nested models test in the context of multiple group models where populations differ from one another (Power simulation). Below, I detail the major findings from this dissertation, how they fit in the current landscape of model evaluation, and their implications.

Major Findings

This simulation serves as a complement to the work of Levy (2017) and demonstrates the importance of taking into account model uncertainty. Specifically, in the Type I Error simulation, it was discovered that the approach of Hancock and Mueller (2011)

that requires two stages of estimation is not suitable for evaluating structural model fit. On the other hand, the approach of Lance et al. (2016) was found to perform well when the estimation model was correct in the population.

A Tale of Two Approaches. To understand the performance of the SM-MV and SM-LV approaches, we must remember what latent variable models provide. Chiefly, the system of equations results in measurement error free latent variables, where observed variance of indicators is partitioned into variance that is shared among indicators and variance that is unique and is not related to the measured latent variable. The Hancock and Mueller (2011) approach is overly optimistic about the ability of latent variable modeling to correct for measurement error and to produce an error free latent variance-covariance matrix. This is apparent based on this approaches total disregard of the uncertainty in the latent variance-covariance matrix and mean vector. In other words, the SM-MV approach puts great confidence in the model-implied variance-covariance matrix, regardless of the strength of the relationship between the latent variable(s) and manifest variables. When measurement quality is poor, the latent variance-covariance matrix and mean vector are estimated poorly and contain noise. Next, when the model-implied latent variance-covariance matrix and mean vector are used as sample moments in the subsequent path analysis, where the latent variables are treated as manifest variables, a tall order has been placed on the estimation of the path model. In other words, the path model attempts to minimize the distance between its model-implied variance covariance matrix and mean vector to *observed* variance-covariance matrix and mean vector that are not properly estimated. This is why the Hancock and Mueller (2011) approach, regardless of SFI, is unable to verify with any consistency that the candidate model is correct in the population. On the other hand, when measurement quality is high, then the model-implied latent variance-covariance matrix and mean vector are estimated with higher accuracy and, therefore, the task of minimizing the distance between the model-implied variance-covariance matrix and mean vector from the path model is more straightforward.

The Lance et al. (2016) approach requires all modeling to be done in the latent space with no changes made to the measurement model across the null, target, and saturated structural models. As a result, the effect of the measurement model is controlled for while evaluating the structural model. By keeping the system of equations intact, the SFIs from this approach are able to more consistently recognize a model that fits in the population. Another manner in which the SM-LV approach differs from the SM-MV approach is how structural model fit is defined. Lance et al. (2016) operationalize structural model fit as the difference between the worst fitting model and the best fitting model. The worst fitting model is one in which the relationship between exogenous and endogenous latent variables is zero (or orthogonal to one another), while the best fitting model is the correlated factors model (i.e., where the structural model is saturated). From this continuum, C9 indices represent the improvement in structural model fit, whereas, C10 indices represent the distance from perfect fit.

Recall that SM_{null} and $PATH_{null}$ were identical, as were the SM_{target} and $PATH_{target}$. This clearly shows the importance of estimating the full system of equations remaining in the latent space. Simply put, when model-implied moments are extracted from a latent variable model and subsequently used as input in a path analysis, problems will surface. This was evidenced by the lack of effect of design factors on the C9 and C10 SFIs from the Lance et al. (2016) approach, whereas, the effect of design factors on the SFIs from the Hancock and Mueller (2011) approach was alarming, with medium-to-large effects observed for the main effect of measurement quality and model size across all SFIs. Relatedly, the interaction between model size and measurement quality was found to have a small-to-large impact on these SFIs.

The performance of the SM-MV was expected based on previous research (Burt, 1976; Levy, 2017; Skrondal & Kuha, 2012). McNeish and Hancock (2018) alluded to problems with the Hancock and Mueller (2011) approach, while at the same time promoting its use for estimating structural versions of common GFIs (e.g., RMSEA and TLI). Further, McNeish and Hancock (2018) state that their two-stage approach could bolster the performance of the Lance et al. (2016) SFIs. McNeish and Hancock (2018) carefully

report the median values for the SFIs and their respective standard deviation, while omitting mean estimates. Without knowing the mean, skew, and kurtosis of the SFI distribution, it is difficult to determine the merit of the approach and its utility for evaluating structural model fit. With a critical look at the standard deviations presented in McNeish and Hancock (2018), a clear effect of measurement quality on their SFIs can be seen. Specifically, as measurement quality decreases, standard deviations increase. These problems appeared not only when using the SFI versions of GFIs, but also the C9 and C10 SFIs of Lance et al. (2016). The suspected problem with the Hancock and Mueller (2011) approach was confirmed by the Type I Error simulation.

The data generating model for the Type I Error simulation simplifies to a single group model (i.e., measurement and structural models were invariant across groups) and generalizes to that of McNeish and Hancock (2018). Because measurement quality had an extraordinary effect on SFIs estimated using the Hancock and Mueller (2011) approach, the SM-MV approach cannot be recommended. For example, partial η^2 was 0.39 for Mc and estimates ranged from 0.022 to 1.004 (HB offer 0.9 as its cut-off); partial η^2 was 0.65 for RMSEA, and estimates ranged from 0.000 to 0.833 (HB offer 0.06 as its cut-off); partial η^2 was 0.36 for CFI, and its estimates ranged from 0.295 to 1.00; and, finally, partial η^2 was 0.08 for χ^2 and ranged from 6.85 to 15,293.21, with the expected value being 22.

The results from the Power simulation echoed previous research that states that structural measures should be preferred over global measures to detect structural model misfit (Lance et al., 2016; McDonald & Ho, 2002). Agreement with prior research (Heene et al., 2011; McNeish & Hancock, 2018) was reached with respect to the impact of measurement quality when the covariance structure is misspecified. By misspecifying the mean and covariance structures independently, it was found that measurement quality had less of an impact on power rates when the mean structure was misspecified. Further, it was found that measures of fit were more sensitive to a misspecified mean structure compared with the covariance structure. When both the mean and the covariance structure were misspecified, measures possessed more power to detect the

model misfit, as expected.

Implications

This study ultimately indicates the use of structural measures of fit that result from single-stage estimation. On the other hand, the Hancock and Mueller (2011) approach, which requires two stages of estimation, should not be used in any fashion. Researchers should be cautious of the SRMR, CFI, and TLI when evaluating structural model fit, as these indices were found to be influenced by measurement quality. The C9 and C10 SFIs, the RMSEA-P, and the $\triangle\chi^2$ measures were found to be reasonable choices for detecting structural model misfit. If the goal is to detect group differences on mean and covariance parameters simultaneously, the suggested structural measures were generally well powered to reject the misspecified model. When the goal is to detect group differences on a single mean parameter that corresponds to either a medium or large effect, nearly all measures of fit were adequately powered to correctly reject the model. When the goal is to detect group differences on a single latent regression, measures are generally well powered to correctly reject the model when the severity of the misspecification is large.

Future Research

Although this study makes unique contributions to the field regarding structural model evaluation, there is much to be investigated. For example, Cole and Preacher (2014) investigated the impact of measurement quality in the context of path analysis. Specifically, they systematically tested the effect of reliability on various variables in the system of equations (e.g., the independent variable, the mediator, or the outcome) to see what impact this had. This is an area that should be investigated in structural equation models, where construct reliability of the exogenous and endogenous LVs are varied. Another fruitful area of research would be investigation of the proportion of high quality measures needed for an LV to be less sensitive to overall measurement quality.

Limitations

As only one model was considered, results may not generalize to other types of models (e.g., difference score models, longitudinal panel models, and formative models) that were not examined. The model utilized in this study was one that was neither overly simplistic nor was overly complex. This decision was made in an effort to generalize the findings to a wide audience.

With respect to data, the simulated data conformed to multivariate normality and was generated without any missingness; therefore, the effect of non-normality and/or missing data on the SM-MV, SM-LV, $\triangle\chi^2$, and RMSEA-P is unknown, based on this study. Another limitation of this study concerns the choice of the baseline model. Much debate surrounds the appropriate way to model the mean structure in the baseline model, and only one baseline model was utilized in the simulations; therefore results may not generalize to situations in which an alternative specification of the mean structure is used. Due to the focus of this study, the constructs were simulated to be fully invariant (e.g., strict invariance) across groups which may not be the most likely scenario for researchers. That said, it was important to have a clean investigation of the sampling distribution and power rates of the measures of fit.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, *103*(3), 411.

Anderson, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on fornell and yi. *Sociological Methods & Research*, *20*(3), 321–333.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303–316.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). Springer.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological methods & research*, *5*(1), 3–52.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological bulletin*, *105*(3), 456.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling*, *14*(3), 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233–255.

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

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(2), 119–143.

Fan, X., & Sivo, S. A. (2009). Using δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, *16*(1), 54–69.

French, B. F., & Finch, H. (2016). Factorial invariance testing under different levels of partial loading invariance within a multiple group confirmatory factor analysis model. *Journal of Modern Applied Statistical Methods*, *15*(1), 26.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological methods*, *16*(3), 319.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, *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.

James, L., Mulaik, S., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data.

Jöreskog, K. G., & Sörbom, D. (1981). *Lisrel v: Analysis of linear structural relationships by maximum likelihood and least squares methods*. University of Uppsala, Department of Statistics.

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, *76*(4), 533–561.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of rmsea in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507.

Lai, M. H., & Yoon, M. (2015). A modified comparative fit index for factorial invariance studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 236–248.

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological methods*, *21*(3), 388.

Levy, R. (2017). Distinguishing outcomes from indicators via bayesian modeling. *Psychological methods*, *22*(4), 632.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International journal of behavioral development*, *31*(4), 357–365.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological bulletin*, *100*(1), 107.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological methods*, *11*(1), 19.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of classification*, *6*(1), 97–103.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods*, *7*(1), 64.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of personality assessment*, *100*(1), 43–52.

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).

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of applied psychology*, *93*(3), 568.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*(5), 869–874.

Moshagen, M., & Auerswald, M. (2017). On congruence and incongruence of measures of fit in structural equation modeling.

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 54–60.

Navarro, D. (2015). Learning statistics with r: A tutorial for psychology students and other beginners. (version 0.5) [Computer software manual]. Adelaide, Australia. Retrieved from `http://ua.edu.au/ccs/teaching/lsr` (R package version 0.5)

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical assessment, research & evaluation*, *9*(11), 8.

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2016). simsem: Simulated structural equation modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=simsem` (R package version 0.5-13)

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from `https://CRAN.R-project.org/package=psych` (R package version 1.8.4)

Rigdon, E. E. (1998). The equal correlation baseline model for comparative fit assessment in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 63–77.

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

Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, 1073191117711020.

Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika*, *77*(4), 649–669.

Steiger, J. H., & Lind, J. C. (May, 1980). Statistically based tests for the number of common factors. In *the annual meeting of the psychometric society. iowa city, ia. 1980.*

Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, *37*(4), 240–256.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. *Structural equation modeling: A second course*, 119–169.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika*, *36*(4), 427–436.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological methods*, *8*(1), 16.

Williams, L. J., & Holahan, P. J. (1994). Parsimony-based fit indices for multiple-indicator models: Do they work? *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(2), 161–189.

Williams, L. J., & O'Boyle Jr, E. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods*, *14*(2), 350–369.

Wu, W., & West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behavioral Research*, *45*(3), 420–452.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from sem and mlm frameworks. *Psychological methods*, *14*(3), 183.

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate behavioral research*, *40*(1), 115–148.

Appendix A

Mean Misspecification

Table A1

Mean Misspecified - Global Measures of Fit

	myMc	myTLI	target.csq	rmsea.afi
pF (A)			0.234	
MQ (B)	0.536	0.572	0.242	0.427
ref.n (C)	0.050	0.041	0.019	0.040
severity (D)	0.868	0.671	0.875	0.822
A:B			0.380	
A:C			0.133	
B:C	0.012	0.034	0.279	
A:D			0.133	
B:D	0.400	0.405	0.471	0.176
C:D	0.032	0.038	0.437	0.016
A:B:C		0.013	0.235	
A:B:D		0.013	0.235	
A:C:D		0.014	0.434	
B:C:D	0.013	0.042	0.607	
A:B:C:D		0.027	0.605	

Table A2

Mean Misspecified - Structural Measures of Fit

	C9.ncp	C9.perDF	C10.ncp	C10.perDF	delta.csq	rmsea.path
pF (A)						
MQ (B)	0.233	0.221	0.233	0.221	0.707	0.601
ref.n (C)	0.093	0.093	0.093	0.093	0.108	0.063
severity (D)	0.907	0.905	0.907	0.905	0.931	0.920
A:B						
A:C						
B:C					0.023	
A:D						
B:D	0.158	0.147	0.158	0.147	0.590	0.242
C:D	0.058	0.057	0.058	0.057	0.064	0.014
A:B:C						
A:B:D						
A:C:D						
B:C:D					0.020	
A:B:C:D						

Appendix B

Variance Misspecification

Table B1

Variance Misspecified - Global Measures of Fit

	myMc	myTLI	target.csq	rmsea.afi
pF (A)			0.666	
MQ (B)	0.723	0.034	0.763	0.624
ref.n (C)	0.046		0.177	0.014
severity (D)	0.666	0.274	0.532	0.606
A:B			0.524	
A:C			0.140	
B:C	0.041		0.256	
A:D			0.532	
B:D	0.486	0.020	0.620	0.261
C:D	0.020		0.241	
A:B:C			0.251	
A:B:D			0.440	
A:C:D			0.230	
B:C:D	0.032		0.840	0.012
A:B:C:D	0.012		0.832	0.014

Table B2

Variance Misspecified - Structural Measures of Fit

	C9.ncp	C9.perDF	C10.ncp	C10.perDF	delta.csq	rmsea.path
pF (A)						
MQ (B)	0.494	0.487	0.494	0.487	0.868	0.822
ref.n (C)	0.060	0.058	0.060	0.058	0.118	0.055
severity (D)	0.783	0.779	0.783	0.779	0.838	0.803
A:B						
A:C						
B:C	0.021	0.018	0.021	0.018	0.100	0.025
A:D						
B:D	0.278	0.272	0.278	0.272	0.715	0.386
C:D	0.024	0.023	0.024	0.023	0.052	
A:B:C						
A:B:D						
A:C:D						
B:C:D	0.018	0.015	0.018	0.015	0.060	0.025
A:B:C:D					0.019	0.016

Appendix C

Mean & Variance Misspecification

Table C1

Simultaneous Misspecification - Global Measures of Fit

	myMc	myTLI	target.csq	rmsea.afi
pF (A)	0.257	0.328	0.970	0.621
MQ (B)	0.797	0.616	0.797	0.739
ref.n (C)	0.113	0.031	0.116	0.067
sev.mean (D)	0.804	0.611	0.804	0.760
sev.var (E)	0.592	0.212	0.593	0.503
A:B		0.081		0.204
B:C	0.031		0.034	
A:D	0.021	0.149	0.024	0.163
B:D	0.249	0.381	0.271	0.020
C:D	0.023		0.026	
A:E	0.035	0.013	0.037	0.025
B:E	0.347	0.017	0.359	0.168
C:E	0.013		0.014	
D:E				0.071
A:B:D		0.056		
A:B:E				0.014
B:D:E				0.010

Table C2

Simultaneous Misspecification - Structural Measures of Fit

	C9.ncp	C9.perDF	C10.ncp	C10.perDF	delta.csq	rmsea.path
pF (A)	0.207	0.132	0.207	0.132	0.363	0.355
MQ (B)	0.012	0.013	0.012	0.013	0.893	0.884
ref.n (C)	0.181	0.179	0.181	0.179	0.225	0.177
sev.mean (D)	0.875	0.874	0.875	0.874	0.897	0.897
sev.var (E)	0.681	0.678	0.681	0.678	0.755	0.748
A:B	0.022	0.022	0.022	0.022	0.015	0.015
B:C					0.066	0.020
A:D	0.013		0.013		0.051	
B:D	0.159	0.157	0.159	0.157	0.442	0.048
C:D	0.054	0.053	0.054	0.053	0.052	0.013
A:E	0.044	0.033	0.044	0.033	0.073	0.036
B:E	0.112	0.112	0.112	0.112	0.542	0.342
C:E	0.013	0.013	0.013	0.013	0.030	
D:E						0.182
B:C:E					0.017	
B:D:E						0.015