ABSTRACT

Title of Dissertation:	CHARACTERIZING UNSYSTEMATIC DISPERSION OF MISFIT IN STRUCTURAL EQUATION MODELS
	Christian T. Meyer, Doctor of Philosophy, Anticipated 2025
Dissertation directed by:	Professor Gregory R. Hancock Program in Quantitative Methodology: Measurement and Statistics Department of Human Development and Quantitative Methodology

Structural equation modeling relies on fit evaluation to assess how well models approximate theoretical relationships, with the goal of providing evidence for the validity of constituent hypotheses. While global fit indices offer convenient summaries of overall model adequacy (Jackson, 2009; McNeish & Wolf, 2023), they cannot guarantee the validity of individual hypotheses, potentially masking substantial localized misspecifications. This limitation becomes particularly problematic in larger models, where acceptable global fit may conceal severe localized misspecification. Current practice supplements global fit with local fit evaluation, but this approach lacks a formalized tolerance for approximation error, unlike global indices that explicitly define acceptable thresholds (Browne & Cudeck, 1993; Hu & Bentler, 1999; Steiger, 2016; Steiger & Lind, 1980).

To address this, the current study makes three key contributions to SEM methodology. First, it introduces misfit dispersion as a new dimension of model fit evaluation. Second, it operationalizes the concept of evenly dispersed misfit through analysis of the statistical properties of local fit, yielding the misfit proportion test for assessing disproportionate local misfit. Third, through comprehensive simulation studies, it evaluates the empirical false positive rate and power of the misfit proportion test in finite samples, establishing guidelines for use in practice.

The novel framework presented here extends the principles of approximate global fit to local fit evaluation, unifying them within a consistent paradigm. By considering both the size of global misfit and its dispersion, this refined criteria better ensure that models and their constituent hypotheses provide acceptable approximations of the substantive processes they represent.

CHARACTERIZING UNSYSTEMATIC DISPERSION OF MISFIT IN STRUCTURAL EQUATION MODELS

by

Christian T. Meyer

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Anticipated 2025

Advisory Committee:
Professor Gregory R. Hancock, Chair & Advisor
Professor Jeffrey R. Harring
Professor Laura M. Stapleton
Professor Peter M. Steiner
Professor Andrea Chronis-Tuscano, Dean's Representative

Dedication

To my family, for helping to shape my humility, empathy, and humor. To Alex, my twin, whose passionate curiosity and determination inspire my own. To my parents, for the stability and guidance you provided through life's unexpected turns.

To the Novick family, for your generous hospitality. To Diane, for your attentiveness and warmth, which fostered a caring environment and filled our days with joy. To Peter, for your invaluable wisdom and limitless capacity for wit.

To Dani – my inspiration, my support, and my best friend, for your embodiment of unwavering dedication to your patients, friends, and family, your reliable presence through every challenge, and your unconditional love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADF	Asymptotically Distribution-Free
CFA	Confirmatory Factor Analysis
CI	Confidence Interval
CRMR	Correlation Root Mean Square Residual
LM	Lagrange Multiplier
LRT	Likelihood Ratio Test
LVPM	Latent Variable Path Model
MI	Modification Index
NT-ML	Normal-Theory Maximum Likelihood
REPCFA	REPeated measures Confirmatory Factor Model
RMSEA	Root Mean Square Error of Approximation
RMSEA_D	Difference Root Mean Square Error of Approximation
SEM	Structural Equation Model(ing)
SRMR	Standardized Root Mean Square Residual

Chapter 1: Introduction

In structural equation modeling (SEM), researchers use fit evaluation to assess how well their model approximates the theoretical relations they aim to study. The essential function is to provide evidence for or against the validity of the model's constituent hypotheses. However, given the complexity of the processes studied with SEM, models are bound to be incomplete or outright incorrect (Meehl, 1990). Practitioners thus strive for models that approximate validity to the extent that they offer a useful basis for "description, prediction, and synthesis" (Cudeck & Henly, 1991, p. 512). Researchers use approximate fit indices (Browne & Cudeck, 1993; Hu & Bentler, 1999; McNeish & Wolf, 2023; Steiger, 2016; Steiger & Lind, 1980) in practice to evaluate the model's fit to a sample as a proxy for its approximation of the population relations.

The primary focus on global fit stems from the convenience it provides by summarizing the combined accuracy of all constituent hypotheses in one measure (Jackson, 2009; McNeish & Wolf, 2023). However, acceptable global cannot imply the validity of any individual hypothesis or subset of them. Like an average, it represents an aggregate of many sources of misfit and thus has the potential to obscure extremes arising from substantial misspecification. Currently, researchers mitigate this limitation by complementing global fit assessment with local fit evaluation, which measures misfit associated with parts of the model rather than the whole.

To make this more concrete, consider the following example. Suppose a researcher concluded that their model with ten degrees of freedom had close fit to a sample with 401 observations. Specifically, the root mean square error of approximation (RMSEA; Steiger, 2016; Steiger & Lind, 1980), a commonly used global fit index (Jackson, 2009), had a 90% upper confidence bound less than 0.05 (Browne & Cudeck; 1993). By inspecting the modification indices (MI; Saris et al., 1987), which measure the misfit associated with individual restrictive

hypotheses in the model, the researcher would find that this limit on global misfit manifests in a limit on local misfit as well.

In the worst case, all global misfit would concentrate into a single misspecified constraint. Yet even then, the maximum possible MI could be at most nine. Although this value is large enough to prompt respecification, SEM practitioners have likely encountered, and ultimately accepted, models with larger MIs. On the other hand, if a model with 120 degrees of freedom achieved similarly close fit, the largest MI could be nearly 200, which would be clearly unacceptable. Without local fit evaluation, researchers cannot rule out this possibility.

At the same time, this example reveals an inherent limitation to local fit evaluation. Whereas global fit indices (e.g. RMSEA) explicitly define an acceptable amount of approximation error, no equivalent framework exists for evaluating local misfit. If we permit approximation error in the aggregate, we should logically extend that tolerance to its parts. Current thresholds for local fit measures, however, derive from the unrealistic assumption of perfect model fit. Adapting local fit evaluation to accommodate approximation error would facilitate a new unified framework for evaluating both closeness of fit and concentration of misfit. This refined approach to defining acceptable approximation would more effectively align fit assessment with its fundamental purpose of providing validity evidence for the hypotheses that constitute a model.

Purpose of Study

The purpose of the current study is to present a novel approach to defining goodness of fit that encompasses both local and global fit. The work draws from the theory of global approximation error and asymptotic analysis of the SEM estimator. Based on this foundation, the study will develop a theoretical characterization of the ideal dispersion of misfit over the model,

termed *evenly dispersed misfit*, and a methodological framework for testing its concentration in a selected subset of the model hypotheses.

This work had the following specific aims:

1) Propose a conceptualization of acceptable misspecification, grounded in the literature on approximate fit.
2) Operationalize evenly dispersed approximation error by extending the statistical properties of local fit measures in models with exact fit to those with approximate fit.
3) Discuss the implications and limitations of the assumptions that underlie this conceptualization and operationalization.
4) Develop a statistical framework for testing the hypothesis of evenly dispersed misfit.
5) Evaluate the test's finite sample properties across diverse models and conditions, specifically investigating:
    a. Differences between the theoretical and empirical false positive rate and power for the test.
    b. Factors influencing the minimum sufficient sample size required for convergence of the empirical distributions of the test to their theoretical asymptotic distributions, including:
        i. Model form and size
        ii. Characteristics and number of hypotheses in the tested subset
        iii. Magnitude of approximation error
6) Discuss the limitations of the proposed statistical framework.

Significance of Study

The current study aims to advance the extant methodological literature studying model fit in SEM by introducing a novel dimension of fit evaluation: the dispersion of misfit. This innovation simultaneously integrates global and local fit evaluation, providing a more comprehensive assessment of model fit. I propose and numerically evaluate statistics and indices for measuring the dispersion of misfit within SEM models. These quantities evaluate the dispersion of misfit alongside global fit and can be naturally incorporated into standard approaches to fit assessment. Furthermore, they can be computed and interpreted as easily as any other statistic provided by SEM software packages, making them accessible and practical for researchers. The approach presented in the study that follows facilitates more precise conclusions about model fit, ensuring that the model, as well as its constituent hypotheses, are acceptable approximations of the substantive processes they represent.

Chapter 2: Literature Review

The structural equation modeling (SEM) methodological literature is saturated with discussions regarding the best ways to evaluate model fit. In concept, fit assessment serves as a gauge of the extent to which the constituent hypotheses of a model hold true within the theoretical population (Maydeu-Olivares, 2017; Mueller & Hancock, 2008). Imperfect fit signals potential issues with model specification; and the degree of misfit provides valuable insights. The poorer the fit, the greater the evidence that a larger number of hypotheses may be incorrect, or that a few misspecified hypotheses are grossly misrepresentative of reality, or some combination thereof.

However, even this idealized conceptualization of model fit presents a major issue: it reduces the multidimensional nature of misspecification to a unidimensional scale. Furthermore, in practice, fit is measured relative to a sample from the population. In this context, the natural desire for a model with perfect fit is unrealistic, given the complex realities being modeled, and even undesirable to avoid overfitting the model to the sample. In the following literature review, I summarize the scholarly efforts to formalize fit evaluation for SEM, recapping the arguments supporting the shift to approximate fit from exact fit, elucidating the tools developed for its practice, examining the shortcomings associated with these tools, and surveying previous endeavors aimed at addressing these limitations.

Foundations of Fit Assessment in SEM

Before discussing how data-model fit is used to evaluate the validity of models in SEM, we must first understand how fit is defined and what causes misfit. In SEM, a model is represented by a parameterized set of equations specifying the relations among observed variables $\mathbf{y}$ (and, potentially, latent variables $\boldsymbol{\eta}$) as well as constraints on the parameters

themselves. These equations define the means $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariances $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ (collectively, moments $\boldsymbol{\beta}(\boldsymbol{\theta}) = \left[\boldsymbol{\mu}(\boldsymbol{\theta})', \text{vech}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)'\right]'$) of the modeled variables and constraints $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ with respect to functions of structural parameters $\boldsymbol{\theta}$.

Models are fit to a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of size n to estimate the true values of the parameters $\boldsymbol{\theta}_0$ and, in turn, provide evidence for or against the validity of the model hypotheses in the population. The logic is as follows: if there are no parameter values $\boldsymbol{\theta}$ satisfying the model constraints that yield, or imply, moments $\boldsymbol{\beta}(\boldsymbol{\theta})$ which are *sufficiently close* to the sample moments $\hat{\mathbf{b}} = [\bar{\mathbf{y}}', \text{vech}(\mathbf{S})']'$, then the hypotheses that constitute the model must be invalid. To this end, the multidimensional differences between $\hat{\mathbf{b}}$ and $\boldsymbol{\beta}(\boldsymbol{\theta})$ are summarized into a univariate measure by a discrepancy function F and, thus, the parameter estimates $\widehat{\boldsymbol{\theta}}$ are defined such that $F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)\right)$ is minimal with respect to all valid values of $\boldsymbol{\theta}$.

It is this concept of discrepancy that has defined and been generalized into the notion of data-model fit used in SEM practice. All SEM fit indices depend, in part, on some measure of discrepancy between the sample and model-implied moments (Hu & Bentler, 1999; Jackson, 2009). When discrepancy is large, it suggests that the fit between the model and sample is poor. Thus, it is reasonable to conclude that the population parameters must be inconsistent with the model hypotheses.

However, the inverse is not necessarily true, as much as we wish it would be. Although small discrepancy, and hence good fit, is desirable, it is not sufficient to imply that the model is valid. There are infinitely many parameter values that would yield the same model-implied moments and discrepancy with the sample moments (Luijben, 1991; Shapiro, 1986). Even if a model could match the population moments exactly, it may still be incorrect. Good fit can, at

best, bolster the theoretical justifications that underlie the model specification. Only poor fit is informative of a model's validity. Paradoxically, this implies that when using fit evaluation to infer model validity, a model that will have some misfit with any sample is preferable to one that can fit all samples perfectly.

This notion is tied to an important topic in SEM known as *identifiability* (Bollen, 1989). A model is said to be *identified* when the parameters are sufficiently constrained such that model-implied moments are distinct for all parameter values that satisfy the constraints. A *just identified*, or *saturated*, model is one that has the same number of parameters as there are moments being modeled. In most cases, this implies that the model maps unique parameter values to all possible moments. Because of this, such models imply moments that exactly match the moments of the sample. Thus, just identified models have perfect fit. *Under identified* models share this property but are insufficiently constrained to yield unique parameter estimates.

When a model has more constraints than are needed for identification, then the model is said to be *overidentified* (Bollen, 1989). Unlike just identified models, the implied moments of overidentified models are restricted to a lower-dimensional subset which corresponds to parameter values that satisfy the model constraints. Consequently, the moments estimated from overidentified models will almost surely have imperfect fit, with implied moments differing from the sample moments.

Overidentification is widely perceived as necessary for assessing model validity (Bollen, 1989). The perfect fit of an under or just identified model provides no evidence toward the validity of the model as it would fit just as well to any sample. Despite sharing the challenges posed by model equivalency as just identified models (Luijben, 1991), the misfit of an overidentified model can be used as evidence against the validity of all models equivalent to it.

The worse the fit to the sample, the more likely it is that the model, as well as all models equivalent to it, is an incorrect representation of the population. In the following section, I will review the literature discussing how the amount of misfit should be measured and interpreted.

Exact fit evaluation

In developing the foundations of SEM, Jöreskog (1969) proposed a statistical test for evaluating goodness of fit. The null hypothesis for the test, denoted by Browne and Cudeck (1993) as the hypothesis of *exact fit*, supposes that there exist parameter values $\widetilde{\boldsymbol{\theta}}$ for which the implied moments $\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ equal the population moments $\boldsymbol{\beta}_0$. When F satisfies the regularity conditions outlined in Shapiro (1986), then the hypothesis of exact fit is tested with the statistic $T = (n-1)F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)$. The discrepancy function in the original proposal by Jöreskog (1969), which was derived from the normal-theory maximum likelihood estimator, is one such example.

Given these conditions on F, the null hypothesis can be re-expressed in terms of discrepancy as $F_0 = F\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})\right) = 0$ and the alternative hypothesis as $F_0 > 0$ (Browne & Cudeck, 1993). Even when the null hypothesis is true, the sample moments $\hat{\mathbf{b}}$ differ randomly from $\boldsymbol{\beta}_0$. So, T for overidentified models is expected to be greater than zero, but there is a limit on how large we should expect it to be. Under the null hypothesis and assumptions about the distribution of the data (Yuan & Bentler, 2006), T asymptotically follows a central χ^2 distribution with $K = p^* - t$ degrees of freedom, where p^* the number of modeled moments for p variables and t is the number of parameters (Jöreskog, 1969; Satorra & Bentler, 1994; Shapiro, 1986; Yuan & Bentler, 1999).

Some methodologists have argued (Barrett, 2007) that the clearly defined and intuitive logic of the statistical test for exact fit makes it the only justifiable way to distinguish good and bad models. Others take issue with the lack of tolerance for any error due to approximation (Browne & Cudeck, 1993; Cudeck & Henly, 1991; MacCallum, 2003; Steiger, 2007). Given the complexity of the processes that are typically modeled in SEM, it is reasonable to assume *a priori* that all are misspecified in some way. Rejecting imperfect models in the pursuit of exact fit runs the risk of overlooking adequate models that nonetheless provide a useful approximation of reality.

Making matters worse, the power of the test of exact fit to reject even trivially misspecified models increases with sample size (MacCallum et al., 1996). This encourages the use of small samples to potentially obscure substantial approximation error within the variability of sampling error (Bentler & Bonett, 1980). Small sample sizes cause the true distribution of T to differ significantly from the asymptotic chi-square distribution, potentially leading to Type I error (Yuan, 2005). Additionally, exact fit rewards overfitting the sampling error with more complex models that have less correspondence to substantive theory but fit better than simpler, more theoretically justifiable models due to having fewer degrees of freedom (Browne & Cudeck, 1993). To address the problems arising from the hypothesis of exact fit, researchers proposed prioritizing a hypothesis of *approximate fit* instead (Browne & Cudeck, 1993; Cudeck & Henly, 1991).

Approximate fit evaluation

The notion of approximate fit follows from the commonly accepted belief that models are unavoidably simpler than the real processes they hypothetically represent (MacCallum, 2003). As noted by Tukey (1961), statistical models are imperfect translations of conceptual models in

the minds of researchers, which are themselves imperfect conjectures about reality. MacCallum (2003), in paraphrasing Tucker et al. (1969), characterized these imperfections in exploratory factor analysis, a special case of SEM, as "the presence and influence of many small common factors, far too many and far too minor to be retained in a factor analysis of empirical data" (p. 135). More generally, Thurstone (1930) argued that "there is probably no law in science which is not easily violated by introducing any of the practical irrelevancies in which the phenomena are actually experienced" (p. 469). From this we must conclude that the best models, in a practical sense, are those that afford "description, prediction, or synthesis" that is meaningful, yet only approximately true (Cudeck & Henly, 1991, p. 512).

For such models in SEM more generally, these "practical irrelevancies" almost certainly violate the hypothesis of exact fit *a priori* (Browne & Cudeck, 1993). In specific terms, if it were possible to fit the model to the population moments $\boldsymbol{\beta}_0$, the resulting parameter estimate $\widetilde{\boldsymbol{\theta}}$ would imply moments $\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ that differ from them. Hence, the discrepancy $F_0 = F\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})\right) > 0$, called the *error due to approximation* by Cudeck and Henly (1991), is nonzero and the hypothesis of exact fit is false. In contrast to exact fit, the framework of approximate fit evaluation allows for approximation error but aims to limit it in some way.

To facilitate approximate fit evaluation, fit indices were developed as alternatives to T. Whereas the value of the fit statistic T depends on several aspects of the model and sample, many fit indices behave as standardized effect sizes of fit, developed with the intention of supporting consistent interpretations across a variety of modeling contexts (Hu & Bentler, 1999; Jackson, 2009; Steiger, 2007; West et al., 2023). Fit indices fall into two categories: nonincremental and incremental (Mueller & Hancock, 2008). Whereas nonincremental fit indices are simple transformations of discrepancy (Mueller & Hancock, 2008; West, 2023), incremental

fit indices are more complex, comparing the discrepancy of the hypothetical model to a baseline model (Mueller & Hancock, 2008; West et al., 2023). Often the baseline model is defined to be the *null model* wherein all variables are independent of each other – at most, means and variances are estimated. The inclusion of the baseline fit measure in addition to the fit of the hypothetical model, typically in a ratio, makes analyzing incremental fit indices mathematically substantially more complicated than nonincremental fit indices. As a result, I will focus most of this literature review on the wealth of studies evaluating nonincremental fit indices.

The two most common indices (Jackson, 2009) are the standardized root mean residual (SRMR) and the root mean square error of approximation (RMSEA). As the name implies, SRMR defines its discrepancy function in terms of the model residuals, which, in the context of SEM, refers to the difference between the sample and model-implied moments $\hat{\mathbf{b}} - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$ (Asparouhov, T., & Muthén, B., 2018; Maydeu-Olivares, 2017). It is computed as

$$\text{SRMR} = \sqrt{\frac{\left(\hat{\mathbf{b}} - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)' \mathbf{W}_{\text{SRMR}} \left(\hat{\mathbf{b}} - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)}{p^*}},$$

where p^* is the number of modeled moments and $\mathbf{W}_{\text{SRMR}}$ is the positive definite diagonal weight matrix that standardizes the moment residuals relative to the variances of the variables. From this definition, SRMR can be understood roughly as the size of the average standardized moment residual for the model. Given the similarity of standardized moment residuals to the familiar effect sizes Cohen's d and Pearson's correlation coefficient, an SRMR greater than 0.1 can be interpreted as a practically significant amount of misfit (Maydeu-Olivares, 2017).

RMSEA was derived with the goal of providing a standardized effect size of the approximation error for the model (Browne & Cudeck, 1993; Steiger, 2016; Steiger & Lind, 1980). Recall that when a model has exact fit with the population, the sample moments will

deviate from the model-implied randomly moments because of sampling error. Under this condition, the fit statistic T asymptotically follows a central a chi-square distribution. However, when the model is misspecified, approximation error induces a systematic shift in T dependent on the size of F_0. This results in T asymptotically following a noncentral a chi-square distribution (Steiger, Shapiro, & Browne, 1985).

Because it represents a direct manifestation of the approximation error for a model, RMSEA was defined in terms of the noncentrality parameter, which is given by $\delta = (n-1)F_0$. It can be estimated from a model fit to a sample to be $\hat{\delta} = T - K$, where K is the degrees of freedom. To emphasize its dependence on discrepancy, as was done with SRMR, I substitute the definition of $T = (n-1)F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)\right)$ to obtain the formula for RMSEA as follows

$$\text{RMSEA} = \sqrt{\max\left(\frac{\hat{\delta}}{K(n-1)}, 0\right)} = \sqrt{\max\left(\frac{F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)\right)}{K} - \frac{1}{n-1}, 0\right)}.$$

The maximum function ensures that any negative estimates of the noncentrality parameter are clamped to zero to ensure the RMSEA is always defined.

In dividing $F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)\right)$ by ν, Steiger and Lind (1980; Steiger, 2016) aimed to penalize overly complex models, which tend to have more freely-estimated parameters and lower degrees of freedom than simpler, more constrained models. The degrees of freedom effectively enumerate the dimensions along which $\hat{\mathbf{b}}$ and $\boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)$ can differ, given the restrictions on $\boldsymbol{\beta}(\cdot)$ imposed by the model constraints. From this we derive a common interpretation of RMSEA as the average amount of discrepancy contributed by each potential misspecification. Models with fewer degrees of freedom have less sources of approximation error and therefore less discrepancy. Dividing by the degrees of freedom ensures that simpler models are preferred over

more complex models, given the same discrepancy. Based on their experience using RMSEA in practice, Browne and Cudeck (1993) suggested that a value above 0.1 indicates that the model is too poor of an approximation to yield valid inferences.

There is a lot that makes these measures similar. First, they share a common mathematical form – the square root of an average discrepancy measure. Additionally, because they can be interpreted in terms of a single, standardized unit of discrepancy, their values ostensibly convey the same meaning in all modelling contexts. Finally, their scales are, on face, remarkably alike, with a meaningful guideline at the value of 0.1. Indeed, the similarities go further than what we have already discussed. Following from large sample theory (Shapiro, 1986; Yuan and Bentler, 2006), we can approximate $F\left(\hat{\mathbf{b}}, \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)$ and thus the square of the RMSEA as

$$(\text{RMSEA})^2 \approx \frac{\left(\hat{\mathbf{b}} - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)' \mathbf{W}_F \left(\hat{\mathbf{b}} - \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)}{K},$$

where $\mathbf{W}_F$ is the positive definite weight matrix derived from the second-order approximation of F. This differs from the square of the SRMR only in the weight matrix and denominator. As we will discuss later in this literature review, the differences between the indices can be traced back to the differences in these components.

The original conception for fit indices was as standardized effect sizes rather than an inferential statistic (Maydeu-Olivares, 2017; Steiger, 2007). This provided a basis for researchers to argue the validity of their model in terms of its ability to approximate the population. To provide a meaningful scale for their values, authors have proposed benchmarks to distinguish between acceptable and unacceptable fit. Analogous to the thresholds for small, medium, and large effect sizes in correlations and mean differences, these cutoff values were selected

somewhat subjectively, drawing on intuition, association with other standardized metrics, and empirical experience (Browne & Cudeck, 1993; Hu & Bentler, 1995; Maydeu-Olivares, 2017).

In practical application, fit index cutoffs evolved to be treated as inferential thresholds, similar to the statistical significance of the fit statistic (Barrett, 2007; McNeish & Wolf, 2023; Millsap, 2007; Steiger, 2007). Instead of challenging the inclinations of practitioners, methodological research sought to formalize this practice. Toward this end, Hu and Bentler (1999) simulated samples for which a model was correctly specified, mildly misspecified, and grossly misspecified and identified cut off values that were consistently able to distinguish among the three conditions. The benchmarks for the previously highlighted indices were 0.95 for the incremental indices Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and Comparative Fit Index (CFI; Bentler, 1990), 0.06 for RMSEA, and 0.08 for SRMR (Hu & Bentler, 1999). These are still considered by many to be the gold standard today (Jackson. 2009; McNeish & Wolf, 2023; West et al., 2023).

Despite their continued popularity, an increasing body of literature has explored limitations in the applications of fit indices and their cutoffs (Hancock & Mueller, 2011; Maydeu-Olivares, 2017; McNeish et al. 2018; McNeish & Hancock, 2018; McNeish & Wolf, 2023; Savalei, 2012; Savalei et al. 2023). Although still regarded as useful, fit indices have been demonstrated to be imperfect measures of both the size and number of misspecifications by these results. In the following section, I will review these studies and detail the implications their findings have on the interpretation of fit indices.

Problems with Fit Indices

Measuring Misspecification Size

To evaluate fit indices, methodological researchers simulate modeling scenarios in which they directly control the number and size of misspecifications (Hu & Bentler, 1999; McNeish & Wolf, 2023; Savalei, 2012; West et al., 2023). In many studies, a misspecification is defined as a constraint that restricts a parameter to a value that differs from the true value in the conceived population. The size of the misspecification is typically regarded as the size of the misspecified parameter bias – the difference between these true and fixed values. The primary purpose of these studies was to identify the conditions that lead to inconsistencies in the interpretations of fit indices and, ultimately, what causes them to fail at differentiating good and bad models.

In studies where the authors varied the size of a single misspecification fit indices behave as expected, increasing with the nominal parameter bias induced by the misspecified constraint (Browne & Cudeck, 1993; Maydeu-Olivares, 2017; Savalei, 2012; Steiger, 2016). Furthermore, studies found that fit indices were sensitive to the number of misspecifications when the misspecifications affected relatively distinct portions of the moment structure. Both SRMR and RMSEA were shown to increase, indicating worse fit as the number of such orthogonal misspecifications increased. This effect was observed for several types of misspecifications including omitted factor correlations (Hu & Bentler, 1999), cross-loadings (Chen et al., 2008; Hu & Bentler, 1999; Savalei, 2012), and error-covariances (Savalei, 2012). However, these results do not hold generally. The association between fit indices and misspecification size complicated by a property of the model-implied moments and, hence, their discrepancy with the sample moments that we will term *parameter overlap*.

Recall that changing the value of a parameter results in a change in the model-implied moments. The overlap between two parameters refers to the similarity of these changes with respect to the discrepancy function. For example, if we have two parameters with a high degree of overlap, then we would expect to measure relatively little discrepancy between the two sets of moments that would result from changing the value of the first parameter while holding the other constant, and vice versa. Similarly, we would expect to measure more discrepancy when changing both parameters simultaneously than the sum of the discrepancy when changing each parameter separately. Because both the model-implied moments and the discrepancy function are nonlinear in the parameters, it is most useful to approximate the change in discrepancy given unit changes in either parameter using the mixed second-order partial derivative of the discrepancy function evaluated at the population or estimated values of the parameters. This measure of parameter overlap underlies commonly-used quantities associated with the relation between parameters, namely the Fisher information matrix, parameter asymptotic covariance matrix, and modification index covariance matrix (Yuan and Bentler, 2006).

Savalei (2012) demonstrated that, when measured by RMSEA, misfit can *even decrease* as misspecifications are added. In the case described in the study, several misspecifications were iteratively added to the model. The misfit increased to begin with, but eventually enough overlapping misspecifications were added that the misfit disappeared. In more realistic scenarios, the impact of parameter overlap on fit indices is less extreme. The most easily observable manifestation of parameter overlap is in post-hoc model modification, wherein releasing a constraint might resolve more misfit when applied prior to releasing a constraint on an overlapping parameter than it would if released after (MacCallum, 1986; MacCallum et al., 1992).

Additionally, parameter overlap is responsible for the well-documented sensitivity that fit indices have to *incidental parameters*. In studies that demonstrate this phenomenon, researchers varied the values of a freely-estimated parameter while holding constant the size of a misspecification. As the model can simply adjust its estimate to accommodate the changing parameter, it was expected that the misfit would not change as a result. However, many studies observed concomitant changes in the values of some fit indices (Chen et al., 2008; Heene et al., 2011; Lai & Green, 2016; McNeish & Wolf, 2023; Mueller & Hancock, 2008; Saris et al., 2009; Savalei, 2012).

The first study in Saris et al. (2009) provides a clear illustration of the problem, demonstrating the sensitivity of RMSEA to changes in the values of incidental parameters in the context of an observed-variable path model. However, much of the concern in the literature is in the context of factor models. Recall that a latent factor represents a construct which is imperfectly measured by observed variables. Both SRMR and RMSEA have been shown to become less sensitive to misspecifications as the manifest variables become less reliable indicators of the latent variables they measure. This has been observed for misspecified constraints on error-covariances (Browne et al., 2002), factor loadings (Chen et al., 2008; Savalei, 2012), factor covariances (Heene et al., 2011), factor dimensionality (Saris et al., 2009), and structural paths (Hancock & Mueller, 2011; McNeish et al., 2018). Hancock and Mueller (2011) termed the inverse relation between measurement quality and fit the *reliability paradox*.

As Hancock and Mueller (2011) discussed in their paper, the reliability paradox arises directly from the choice of discrepancy function in these fit indices. Whereas discrepancy specifically measures the model's approximation of the observed variable distribution, the misspecifications in these examples were in the relations among the latent variables. Thus, the fit

indices are merely indirect measures of the true misspecifications, confounded by aspects of the observed variables that are entirely incidental.

Taken together, these problems arising from parameter overlap have been taken as evidence of inconsistency in the scales of many fit indices, implying that interpretations of their values in one modeling context do not necessarily generalize to others (McNeish & Wolf, 2023). Hancock and Mueller (2011) attempt to resolve the reliability paradox directly by proposing a new discrepancy function based on the estimated latent variable covariance matrix from a saturated model instead of the measured variable covariance matrix. This effort exemplifies a more general approach to rectifying misalignments between the expectations of misspecification size and how fit indices measure them in practice. To adequately resolve these problems, it may be necessary to define clearly and comprehensively what we believe to be the practically meaningful impact of misspecifications and choose or design fit indices that formally operationalize this definition.

Rather than change the scales themselves, an alternative approach proposed in the literature is to change the benchmark values for fit indices. Millsap et al. (2012) proposed a framework extending the method used in Hu and Bentler (1999) to tailor fit index cutoffs to the modeling context. McNeish and Wolf (2023) improved on that method with their dynamic fit index cutoff method, which automatically identifies cutoff values based on their ability to differentiate simulated samples where the practitioner's hypothesized model is either correctly or incorrectly specified.

Scaling with Model Size

Larger models and those with more degrees of freedom have more ways to be misspecified than smaller models and complex models estimating many parameters relative to

their size. Hence, we would expect the total discrepancy to scale with model size and complexity (Maydeu-Olivares, 2017; Steiger, 2016). By averaging the overall discrepancy over the model size, both RMSEA and SRMR adjust for these factors, in theory. However, in practice, it is possible for the contribution of large misspecifications to be swamped by good fit elsewhere.

Several papers demonstrated the tendency for RMSEA of a model fit to a population to decrease as the model degrees of freedom increase. Three studies noted this trend for the population RMSEA as the size of the model increased, but the number of misspecifications remained fixed (Kenny et al., 2003; Maydeu-Olivares, 2017; Shi et al., 2019). This result was consistent across distinct types of misspecifications – specifically, omitted error-covariances and incorrect factor dimensionality. Additionally, Breivik et al. (2001) considered models with an increasing number of 3-indicator latent factors. Although the number of misspecified factor correlations or cross-loadings increased linearly with the number of factors, the degrees of freedom increased quadratically, quickly washing out the impact of the misspecifications.

In contrast, when applied to a model fit to a sample instead of a population, the sample RMSEA can increase with model size. The estimate for the noncentrality parameter used in the RMSEA discrepancy component is biased in finite samples (Yuan, 2015). Moshagen et al. (2012) found that this bias increased with the model degrees of freedom. Additionally, Shi et al. (2019) demonstrated that this finite-sample bias grows much faster than the decreasing trend in the population RMSEA, even in samples as small as five hundred observations. Notably, this occurred for models with greater than sixty variables that had more than 1750 degrees of freedom, which would be considered large for many applications of SEM.

As sample size decreases, the size of sampling error relative to the discrepancy caused by misspecifications increases. When using a fixed cutoff value for RMSEA to accept or reject a

model, the false positive rate was shown to increase as sample size decreased (Chen et al., 2008). Kenny et al. (2015) noted the effect to be particularly stark for models with low degrees of freedom, and thus cautioned against using RMSEA in these cases.

Studies evaluating SRMR indicate that it is not as susceptible to the model size effect as RMSEA. Shi et al. (2022) found the rejection rates based on SRMR for one factor confirmatory models with low degrees of freedom were acceptable for all model size conditions. Furthermore, Maydeu-Olivares (2017) demonstrated that SRMR and the related correlation root mean square residual (CRMR) were much less sensitive to model size than RMSEA in the context of fitting a one factor model to a population with two correlated factors. However, it is unclear whether these results generalize to other modeling contexts. For example, because SRMR and CRMR treat each standardized moment as an independent source of misfit, misspecifications that impact relatively few moments are more likely to be swamped by the lack of misfit in the other moments. Additional research is needed to clarify the nature and extent of the model size effect for SRMR.

To mitigate the risk of good fit in large models obscuring misspecifications, the literature suggests supplementing global fit evaluation with component-wise or local fit evaluation. Some models can be broken down conceptually into components or submodels that can be separately evaluated for goodness of fit. A common example is the decomposition of the model into a measurement and structural component (Anderson & Gerbing, 1992; Fornell & Yi, 1992; Lance et al., 2016, Rosseel & Loh, 2024). In this approach, the latent variable model is saturated, meaning that, if it were instead an observed variable path model, it would be just identified. None of the misfit left in the model can be attributable to structural misspecifications and thus must be attributable to misspecification in the measurement model. Thus, the fit indices

computed from the discrepancy for the structurally saturated model represents a pure measure of the measurement model misspecification size. Measuring the fit for the structural model is less direct and requires alternative approaches to measuring discrepancy like those presented in Hancock and Mueller (2011) and Lance et al. (2016). Alternatively, Lagrange multiplier-based methods like the modification index and expected parameter change (Saris et al., 1987; Saris et al., 2009; Whittaker et al., 2012) or graphical criteria tests like implied partial correlations and tetrads (Thoemmes et al., 2018) can be used to identify more general misspecifications. However, MacCallum et al. (1992) warned that exploratory evaluation of local fit indices may lead to inferences that do not generalize out of sample. A theory driven approach to modification helps mitigate this problem (MacCallum, 1986).

Although useful at identifying potential problems with a model, it is unclear how these measures work within the approximate fit evaluation framework. Inferences about local fit typically made with respect to the hypothesis of exact fit. Modification indices and nested-model comparison likelihood ratio tests, for example, are compared against the critical values of a central chi-square distribution (Saris et al., 1987). There is, to date, no guidance for how to interpret component-wise (e.g., measurement and structural components) and local fit indices if we consider a certain amount of global approximation error acceptable.

In the following section, I will argue that the criteria for acceptable model fit should be extended to account for the dispersion of local misfit across the hypotheses comprising the model. Using the asymptotic theory of the SEM estimator, I will develop a mathematical framework for quantifying this misfit dispersion, anchored by an ideal benchmark: the even dispersion observed under conditions of exact fit. Building on this foundation, I will propose a

statistical test for evaluating dispersion, which can be seamlessly integrated into standard SEM practice.

Chapter 3: Theoretical Formulation

The preceding literature review underscores the urgent need to establish a rigorous, comprehensive, and broadly agreed-upon definition of *adequate* fit in SEM. To advance the field of model fit evaluation, this definition must reflect widely-held conceptualizations of both the size and the number of misspecifications. It should also be applicable across diverse modeling contexts and be particularly sensitive to the areas of the model that researchers are most interested in drawing inferences from, such as the structural components of a latent variable path model. These considerations are crucial for improving how fit indices measure discrepancies between the sample and model-implied moments. Whether through universally applicable measures or bespoke indices tailored to specific models, these efforts to refine the definition of adequate fit can help ensure that fit evaluation can identify models that are useful insofar as they provide a reasonably acceptable approximation of reality.

That said, focusing solely on these global measures neglects a deeper objective of model evaluation: ensuring the validity of our inferences about the model's constituent hypotheses. The challenge lies in moving beyond merely assessing whether a model, as a whole, fits well to examining how misfit is dispersed across the model. A nuanced approach that evaluates both dimensions simultaneously is essential for achieving a more accurate and meaningful understanding of a model’s performance.

Conceptual Framework for Evenly Dispersed Misfit

Structural equation modeling imposes inherent limitations on the types of relations that can be represented, which extend beyond the restrictions of modeling mean and covariance structures alone (Meehl, 1990). While real-world data-generating processes typically exceed the complexity of specified models, these simplified representations can still yield meaningful

approximations (Cudeck & Henly, 1991; MacCallum, 2003; Thissen, 2001). MacCallum (2003) contended that researchers must accept imperfect models provided they contain no theoretically avoidable misspecifications. This perspective acknowledges the inevitable presence of minor violations stemming from innumerable insignificant population characteristics. An example imagined by Tucker, Koopman, and Linn (1969) described the presence of numerous small common factors that would be too trivial to retain in factor analysis yet contribute to the overall misfit. When the rest of the model is supported by sound theoretical justification, such misspecifications are realistically unfixable.

This acknowledgment does not imply that better-fitting models cannot be specified within existing modeling frameworks. Empirical practice routinely demonstrates that modifications can improve global fit. However, data-driven approaches like Modification Index-guided respecification often identify many meaningless constraint subsets whose release artificially improves fit (MacCallum, 1986; MacCallum et al., 1992). Such modifications risk capitalizing on sampling variability while offering no guarantee of replicable improvement. The principle of parsimony (Preacher, 2006; Preacher & Haley, 2023) suggests that researchers must accept some misspecifications when their resolution would introduce untenable complexity. In summary, methodologists recommend avoiding modifications that may improve fit but ultimately fail to fix meaningful misspecifications and generate more interpretational problems than they resolve.

These considerations yield a well-defined characterization of acceptable model misspecification. After accounting for theoretical justification and interpretability, the approximation error contributed by the remaining fixable misspecifications must be uniformly small. In practical terms, this results in two implications for fit evaluation. First, it aligns with traditional standards requiring limited global approximation error. Second, it establishes that

global goodness of fit is sufficient only if the approximation error attributable to each subset of model hypotheses is limited as a function of the complexity it would add to the model if modified. I term this new condition *evenly dispersed approximation error*.

This conceptual framework requires careful operationalization. At minimum, any definition must remain consistent with the behavior of misfit under exact fit conditions. The following section develops this operationalization through analysis of the statistical properties for the likelihood ratio test of exact fit and extending these properties to approximate fit scenarios.

Operationalizing Misfit Dispersion

Let us consider how misfit is dispersed when a model has exact fit to a population. Recall that, in specifying a model, we define both a model-implied moment function $\boldsymbol{\beta}(\cdot)$, which maps parameters $\boldsymbol{\theta}$ to the moments of the modeled variables, and a set of parameter constraint equations $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ (ignoring inequality constraints for simplicity). Suppose we are interested in evaluating the validity of this model in a population with parameters $\boldsymbol{\theta}_0$ implying moments $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{\theta}_0)$ and we choose a discrepancy function F to measure the deviation of any model-implied moments. A model is said to have exact fit if there are parameter values $\widetilde{\boldsymbol{\theta}}$ that satisfy the model constraints $\mathbf{h}(\widetilde{\boldsymbol{\theta}}) = \mathbf{0}$ and imply moments $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ that perfectly match the population moments $\boldsymbol{\beta}_0$ such that $F\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})\right) = 0$.

Fitting the model to a sample from this population with moments $\boldsymbol{\beta}_n$ yields the parameter estimate

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}}{\operatorname{argmin}} F\big(\boldsymbol{\beta}_n, \boldsymbol{\beta}(\boldsymbol{\theta})\big).$$

The global fit likelihood ratio test (LRT) statistic is given by $T = (n-1)F\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})\right)$. If we modify the model by releasing $1 \leq k < K$ independent constraints on the model parameters, then fitting the modified model to the sample would yield a new estimate $\widehat{\boldsymbol{\theta}}_{mod}$ with global fit statistic $T_{\text{mod}} = (n-1)F\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_*(\widehat{\boldsymbol{\theta}}_{mod})\right)$, where $\boldsymbol{\beta}_*(\cdot)$ is the modification model-implied moment function. The difference statistic $T_{\text{diff}} = T - T_{\text{mod}}$ represents the improvement in global fit following the modification.

Given that the misfit under the condition of exact fit is due to random sampling error, we can characterize the dispersion of these measures in terms of how they are randomly distributed. To do so, we first need to introduce some notation. Let $2\mathbf{V}$ be the Hessian of $f(\boldsymbol{\beta}) = F(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}_0$ and $\boldsymbol{\Delta} = \left[\boldsymbol{\Delta}_f', \boldsymbol{\Delta}_c'\right]'$ be the Jacobian of $\boldsymbol{\beta}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$, where f and c denote the columns corresponding to the free and constrained parameters, respectively. It follows from the central limit theorem that

$$\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{\Gamma}).$$

that T can be approximately expressed as a quadratic form of $\boldsymbol{\beta}_n - \boldsymbol{\beta}_0$,

$$T = (n-1)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)'\mathbf{U}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) + o_{p^*}(1).$$

where the matrix $\mathbf{U} = \mathbf{V} - \mathbf{V}\boldsymbol{\Delta}_f\left(\boldsymbol{\Delta}_f{}'\mathbf{V}\boldsymbol{\Delta}_f\right)^{-1}\boldsymbol{\Delta}_f{}'\mathbf{V}$, large sample theory (Shapiro, 1986; Rao & Mitra, 1971; Yuan and Bentler, 2006) tells us that

$$T \xrightarrow{\mathcal{L}} \chi^2_K,$$

where $K = rank(\mathbf{U}\boldsymbol{\Gamma})$, the model degrees of freedom. To be specific, F and $\boldsymbol{\beta}(\cdot)$ must be defined such that $\mathbf{U}\boldsymbol{\Gamma}\mathbf{U} = \mathbf{U}$ ($\mathbf{U}$ is a generalized inverse of $\boldsymbol{\Gamma}$) or, at least, $\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}$.

From Lagrange multiplier theory, we can derive an approximation of T_{diff} for a given modification with k degrees of freedom without needing to modify and refit the model. Let $\boldsymbol{\Delta}_k$

be the subset of columns of $\boldsymbol{\Delta}_c$ that correspond to the k independent constraints that would be released in the modification. Then, T_{diff} can also be expressed as a quadratic form of $\boldsymbol{\beta}_n - \boldsymbol{\beta}_0$,

$$T_{\text{diff}} = (n-1)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)'\mathbf{U}_k(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) + o_{p^*}(1),$$

where $\mathbf{U}_k = \mathbf{V}\boldsymbol{\Delta}_k(\boldsymbol{\Delta}_k'\mathbf{U}\boldsymbol{\Delta}_k)^{-}\boldsymbol{\Delta}_k'\mathbf{V}$. Following the same logic as before, we have that

$$T_{\text{diff}} \xrightarrow{\mathcal{L}} \chi_k^2.$$

Furthermore, from the theory multivariate distributions (Fang & Zhang, 1990; Rao & Mitra, 1971), we know that the proportion of misfit resolved by the modification T_{diff}/T is independent of T, and has the asymptotic distribution

$$T_{\text{diff}}/T \xrightarrow{\mathcal{L}} Beta\left(\frac{k}{2}, \frac{K-k}{2}\right).$$

Note that because $\mathbf{U}_k$ is invariant to the choice of g-inverse $(\boldsymbol{\Delta}_k'\mathbf{U}\boldsymbol{\Delta}_k)^{-}$, this result holds for any subset of constraints for which $\text{rank}(\boldsymbol{\Delta}_k'\mathbf{U}\boldsymbol{\Delta}_k) = k$, even if the subset contains redundant or linearly dependent constraints (Rao & Mitra, 1971; Shapiro, 1986).

The final result is a defining property of not just the central normal distribution, but all members of the central elliptical family of distributions. Following from the theory of quadratic forms of central elliptical distributions (Fang & Zhang, 1990), we can see that the dispersion of misfit is more even than originally implied above. Suppose we have elliptically distributed variable $\mathbf{x} \sim EC(\mathbf{0}, \boldsymbol{\Sigma}, \phi)$, where the matrix $\boldsymbol{\Sigma}$ describes the ellipticity of $\mathbf{x}$ in the same way $\boldsymbol{\Gamma}$ defined the covariance of $\boldsymbol{\beta}_n$. Then the quadratic form of $\mathbf{x}$ with respect to matrix $\mathbf{U}_k$ that satisfies $\mathbf{U}_k\boldsymbol{\Sigma}\mathbf{U}_k = \mathbf{U}_k$ is distributed as

$$\mathbf{x}'\mathbf{U}_k\mathbf{x} \sim R,$$

if $\text{rank}(\mathbf{U}_k) = K$, and

$$\mathbf{x}'\mathbf{U}\mathbf{x} \sim R\sum_{i=1}^{k} D_i,$$

if $1 \leq \text{rank}(\mathbf{U}_k) = k < K$, where R is an arbitrarily distributed positive scalar that is independent of $\mathbf{D} = (D_1, \ldots, D_K)' \sim Dirichlet\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$. Thus, not only are the proportions T_{diff}/T identically distributed, but they can be expressed as the sum of smaller components of misfit that have the same marginal distribution as the proportion resolved by a modification with 1 degree of freedom.

These properties of misfit under the condition of exact fit bear a noticeable resemblance to the conceptualization of unavoidable approximation error given above. First, the global misfit as measured by T is definitionally small, especially so in the context of approximate fit evaluation. Second, the misfit is evenly dispersed, given that the proportion of misfit resolved by a modification is identically distributed, asymptotically, as any other modification of the same size. Third, from the independence of T and T_{diff}/T, the misfit is evenly dispersed regardless of the amount global misfit. Let us refer to these conditions, collectively, as evenly dispersed misfit.

When the model does not have exact fit to the population, the model is misspecified with respect to the population, such that the approximation manifests as discrepancy between the population and model-implied moments, $F_0 = F\left(\boldsymbol{\beta}_0, \boldsymbol{\beta}\left(\widetilde{\boldsymbol{\theta}}\right)\right) > 0$. As a result, the moment residuals $\boldsymbol{\beta}_n - \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)$ are shifted such that they are asymptotically distributed as

$$\sqrt{n}\left(\boldsymbol{\beta}_n - \boldsymbol{\beta}\left(\widehat{\boldsymbol{\theta}}\right)\right) \xrightarrow{\mathcal{L}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu} = \boldsymbol{\beta}_0 - \boldsymbol{\beta}\left(\widetilde{\boldsymbol{\theta}}\right)$ (Steiger, Shapiro, and Browne, 1985). The asymptotic distribution of T is consequently a noncentral chi-squared, $\chi^2_K(\delta)$, with noncentrality parameter

$$\delta = (n-1)F_0 \approx (n-1)(\boldsymbol{\mu}'\mathbf{U}\boldsymbol{\mu}).$$

Similarly, T_{diff} are asymptotically distributed as $\chi^2_k(\delta\pi_{\text{diff}})$, where $\pi_{\text{diff}} = 1 - F_1/F_0 \approx (\boldsymbol{\mu}'\mathbf{U}_k\boldsymbol{\mu})/(\boldsymbol{\mu}'\mathbf{U}\boldsymbol{\mu})$, where F_1 is the discrepancy for the modified model.

Because $\mathbf{U}_k$ differs between nonredundant constraint subsets, then δ_{diff} is almost surely different for all modifications of the same size, and thus, the T_{diff}/T for modifications of size k are not guaranteed to have the same asymptotic $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ distribution. Instead, they are asymptotically distributed as $Beta\left(\frac{k}{2}, \frac{K-k}{2}; \delta\pi_{\text{diff}}, \delta(1-\pi_{\text{diff}})\right)$, the doubly noncentral beta distribution (Orsi, 2017). Therefore, the addition of approximation error necessarily causes misfit to be dispersed unevenly.

The extent to which the uneven dispersion of misfit under the condition of approximate fit deviates from exact evenness depends on the degree to which the approximation error can be systematically and disproportionately attributed to specific misspecifications to the model. Suppose that approximation error was not fixed, but instead varied randomly, independent of sampling error and distributed elliptically. Because the sum of two elliptical variables is necessarily elliptical (Hult & Lindskog, 2002), the misfit resulting from the combination of sampling and approximation error, the total error, would be elliptical and thus evenly dispersed, despite the fit not being exact. Notably, we do not need to assume that approximation error is random. We only need to ensure that it does not cause the total error to measurably deviate from what would be expected if it were.

Hence, under the condition of evenly dispersed misfit, T_{diff}/T is asymptotically distributed as $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$. This holds true whether the model has exact fit to the population or whether its approximation error is indistinguishable from sampling error. Building upon this theoretical foundation, the following section develops a statistical framework for testing that the

local proportion of misfit, as measured by T_{diff}/T, is consistent with the hypothesis of evenly dispersed approximation error.

Misfit Proportion Test

Given a subset of constraints, the misfit proportion statistic T_{diff}/T can be used to statistically test the hypothesis of evenly dispersed approximation error. If the observed value of the statistic surpasses the $(1-\alpha)$ quantile of the $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ distribution, we can reject the hypothesis, concluding uneven dispersion of approximation error with asymptotic false positive rate α. The power of the test to detect alternatives to evenly dispersed misfit depends on α and the characteristics of the misspecifications. These properties manifest in the noncentrality parameters of the asymptotic alternative distribution, $Beta\left(\frac{k}{2}, \frac{K-k}{2}; \delta\pi_{\text{diff}}, \delta(1-\pi_{\text{diff}})\right)$, which depend on the sample size n, global approximation error F_0, and most saliently, the proportion of approximation error attributable to the hypothesized constraint subset π_{diff}.

Intuitively, the larger π_{diff} is, relative to the proportion of the model degrees of freedom contained in the modification subset k/K, the more disproportionally the approximation error is dispersed, and hence the statistic should have more power to detect it. However, this unevenness can be attenuated, randomly, by the evenly dispersed misfit from sampling error. Because sampling variability increases with sample size n, it can dwarf approximation error when both F_0 and n are relatively small. Conversely, when n is large, approximation error can dominate sampling variability even when F_0 is small. This is a property inherited from the LRT statistics on which the test is based. Whereas T is considered to be overpowered at moderate sample sizes (Bentler & Bonett, 1980; Jöreskog, 1969; Tucker & Lewis, 1973), the additional power gained with moderate sample size may be necessary for T_{diff}/T to be useful. Even when $\pi_{\text{diff}} = 1$, such

that all approximation error is attributable to the hypothesized constraint subset, the misfit proportion test can be underpowered to detect the unevenness if F_0 and n are too small. This limitation will be explored in an asymptotic power analysis performed in the next chapter.

In addition to raising questions about the asymptotic power of the test, this dependence on LRT statistics may have analogous impacts on its performance in finite sample sizes. Assuming a model has exact fit to a normally distributed population, T is only ever approximately distributed as χ^2 (Bentler, 1990; Jöreskog & Sörbom, 2006). The deviation of the true sample distribution from the asymptotic distribution causes the false positive rate to differ from the nominal α at finite sample sizes. This error decreases to insignificance at moderate n, but the speed of convergence depends on the size of the model, slowing as the number of modeled variables and degrees of freedom increase (Moshagen, 2012; Yuan, 2005; Yuan and Bentler, 2006, Yuan et al., 2015). When the model is misspecified, Yuan (2005) demonstrated that the distribution T in finite samples is more closely approximated by a normal distribution than the noncentral χ^2, unless the model is trivially misspecified.

If these finite sample errors are propagated to T_{diff}/T, we can expect differential impacts on the false positive rate, depending on whether the evenly dispersed misfit is due to exact fit or evenly dispersed approximation error. Power would be doubly affected, compounding finite sample errors from both the null and alternative distributions. As such, it is vitally important to evaluate the empirical performance of the misfit proportion test in finite samples across a range of sample sizes, models, and misspecification conditions. In the following chapter, I report the design and results of a sequence of simulation studies aimed at addressing these concerns.

Applications for the Misfit Proportion Test

The misfit proportion test facilitates the evaluation of *approximate local fit* as simple, yet effective alternative to the inherently flawed evaluation of *exact local fit*. Most commonly, local fit is evaluated through the framework of nested model comparisons, typically using the likelihood ratio (LR) difference test, from which the misfit proportion test is derived. Within this framework, the fit of a more constrained parent model is compared to that of a less constrained child model, both estimated from the same sample. The models must be nested for the statistical properties of both the LR difference test and the misfit proportion test to hold. This requires that the parent model can be transformed into the child model solely by releasing constraints. The resulting difference in fit corresponds to the misfit attributable to the distinct constraints between models, effectively measuring the misfit resolved by a potential modification T_{diff}.

For local fit evaluation, nested model comparison serves two key functions. First, it supplements global fit evaluation to strengthen inferences regarding model adequacy. While global fit indices assess overall model accuracy, researchers must demonstrate limited local misfit throughout the model to support the validity of its constituent hypotheses. Second, it aids in identifying specific misspecified hypotheses to inform model respecification.

Given the multidimensional nature of local misfit, an exhaustive test of all constraint subsets would be required to definitively rule out misspecification. However, this is often computationally infeasible. Consequently, the scope must be restricted to a tractable number of subsets, justified theoretically by arguing that untested subsets (including subsets of tested subsets) are unlikely to represent nontrivial misspecifications. Subset selection can be approached in a confirmatory or exploratory manner.

In confirmatory applications, constraint subsets must be defined *a priori*, ideally before collecting the sample, alongside justification for their exhaustiveness. For example, in multigroup invariance testing, a nested sequence compares metric (loadings), scalar (intercepts), and strict (error variances) invariance models. Exploratory approaches, conversely, rely on *post hoc* algorithms (e.g. stepwise variable selection) to select subsets. Justification for the exhaustiveness of the algorithm, especially the selection heuristic and stopping rule, is required only if the researcher aims to infer an absence of misspecification from a null result. Otherwise, purely exploratory investigations seeking to identify sources of misfit need not establish exhaustiveness.

Substituting the misfit proportion test (or similar approximate local fit methods) for the exact fit LR difference test offers several advantages. Conceptually, it resolves the incoherence of supplementing approximate global fit assessment with exact local fit tests. Practically, it almost always imposes a higher threshold for detecting misspecification than exact fit tests. Importantly, this threshold scales with sample size, maintaining consistent sensitivity to a fixed level of approximation error. As a fortuitous side effect, this consequently privileges the *a priori* theoretical specification of the base model over data-driven respecification.

These benefits can be illustrated through forward stepwise specification search. The process traditionally involves: (1) inspecting modification indices (MIs), which are asymptotically equivalent to 1-df LR difference tests for exact fit; (2) releasing the constraint with the largest MI; (3) refitting the model; and (4) iterating until no MI exceeds the exact fit test threshold. This data-driven procedure has been criticized for its susceptibility to capitalization on chance (MacCallum et al., 1992), risk of overfitting sampling error, and reduced parsimony

(Preacher, 2006). As I will demonstrate through the following example, using approximate local fit measures can mitigate these issues.

Consider a model with $K = 100$ degrees of freedom. Assume we have three different populations: one with exact fit ($RMSEA = 0.00$), and two with evenly dispersed approximate error ($RMSEA = 0.05$). The model is fit to samples drawn from each population. The samples from the approximate fit populations had $n = 401$ and 801 observations each. For the exact fit sample, n is left undefined as it is irrelevant. To provide a sense for how these three different conditions affect MIs, let's consider the expected value and 0.95 quantile of an individual MI. Under exact fit, the expected value is 1 and the quantile is 3.84. Under evenly dispersed approximate fit, the expected values increase to 2 and 3, and the quantiles to 7.67 and 11.50, respectively. As we can see, the MI distribution increases with both approximation error and sample size, with the average in the largest sample of the approximate fit condition approaching the 0.95 quantile of the exact fit condition.

Despite all misspecifications being trivial by definition due to evenly dispersed approximation error, the sheer number of MIs (potentially > 200) creates a severe multiple testing problem. The probability of one or more false rejections, called the familywise error rate (FWER), can substantially exceed the nominal false positive rate $\alpha = 0.05$. Controlling FWER requires adjusting the per-test false positive rate (α_c) based on the unknown dependence structure among MIs. Assuming a dependence structure in which 100 independent pairs of perfectly dependent MIs yields $\alpha_c = 1 - (1 - \alpha)\text{^}(1/100) \approx 0.00051$. The FWER would be $1 - (1 - p)\text{^}100$, where p is the probability that an individual test would be rejected, given the approximation error and sample size conditions. Properly applied, this correction controls FWER at 0.05 when the null hypothesis of the test is true.

Recall that the exact fit test compares each MI to the quantiles of the χ_1^2 null distribution. Applying the multiple testing correction yields a rejection threshold of 12.07 for the independent case. If the MIs follow the assumed dependence structure, the FWER would be 0.05 under exact fit because the null hypothesis of the test is true. However, under approximate fit, the FWER was uncontrolled. For $n = 401$, FWER equaled 0.752; and for $n = 801$, it equaled 0.990.

On the other hand, recall that the misfit proportion test computes ratio of the MI to global misfit and compares it to the quantiles of the $Beta(1/2, 99/2)$ distribution. Given the multiple testing correction, the threshold for the test was 0.115. Notably, all conditions satisfy evenly dispersed approximation error, the null hypothesis for the misfit proportion test. Thus, if the MIs had the assumed dependence structure, the FWER was successfully controlled at 0.05 for all samples.

These examples illustrate how the misfit proportion test enhances specification search, when the model has approximate fit. The exact fit test suffered from reduced specificity as MIs inflated with sample size and approximation error, leading to high probabilities of erroneous modification. In contrast, the misfit proportion test maintained consistent specificity. Had the approximation error been unevenly dispersed, the exact fit test would be overpowered even for trivial concentrations. The power of the misfit proportion test scales more sensibly, depending primarily on the disproportionality of the error concentration on the tested constraint. Consequently, respecification guided by the misfit proportion test should reliably terminate after addressing only non-trivial misspecifications, leaving the trivial misspecifications unmodified. Therefore, when respecifying a model with approximate fit, substituting the misfit proportion test for the exact fit LR difference test can help limit the risk of overfitting and capitalization on

chance, retaining more of the theoretically justified specification from the original model by minimizing the reliance on potentially spurious data-driven reasoning.

Chapter 4: Simulation Studies

Because the misfit proportion statistic is derived from two statistically dependent likelihood ratio test (LRT) statistics, important questions remain about its finite sample behavior. These concerns relate to previous empirical findings regarding the convergence of likelihood ratio test statistics (Yuan, 2005) and the well-documented model size effect (Moshagen et al., 2012; Shi et al., 2019), where LRT statistics demonstrate increasing positive bias with larger model complexity. Additionally, it is unclear how the biasing effects of approximation error on the alternative distribution of LRT statistics (Yuan et al., 2007) will propagate to the statistic and interact with the other biases.

To address these concerns, we conducted two simulation studies evaluating the finite sample properties of the misfit proportion test, specifically its empirical false positive rate and statistical power. Both studies manipulated three factors: (1) model degrees of freedom, (2) size of tested hypothesis subsets, and (3) magnitude of approximation error. These studies aimed to answer two critical research questions: First, how do the manipulated factors influence finite sample bias? Second, at what sample size does this bias become practically negligible? These questions were addressed by analyses of the bias and convergence, respectively, in both studies.

For the bias analysis, we regressed the finite sample bias on the manipulated factors to estimate the size and directionality of their effects. For the convergence analysis, we defined a practical significance threshold for bias magnitude, then identified the minimum sample size where bias fell below this threshold. Together, these analyses were meant to inform guidelines for the use of the misfit proportion test in practice. The results of the bias analysis enable practitioners to determine how well they should expect the test to adhere to its asymptotic properties in their own applications, while the results of the convergence analysis are meant to

provide sample size recommendations to ensure that using asymptotic distribution for the test is appropriate.

In the rest of the chapter, we enumerate the manipulated model and sample factors in the simulation studies, detail the methods used to generate the simulated data, and describe the design parameters for each study. Then, we explain in depth the methodology used in the bias and convergence analyses before finally presenting the results.

Simulation Data

Simulated data were generated relative to three hypothetical models: a three-factor confirmatory factor analysis (CFA), a five-factor latent variable path model (LVPM), and a repeated measures version of the CFA (REPCFA) at two time points. For each model, I evaluated the performance of three misfit proportion statistics corresponding to distinct constraint subsets. These models and subsets were chosen to represent a range of forms, sizes, and types of constraints. The variety of model forms was limited based on previous studies of the LR difference test (Moshagen et al., 2012; Shi et al., 2019) to focus on conditions in which convergence was achievable at moderate sample sizes ($n = 200$).

In Moshagen et al. (2012), the empirical false positive rate had greater than 40% relative bias, compared to nominal level, at $n = 200$ for a model with $K \approx 80$. At the next largest model size ($K \approx 400$) analyzed in the paper, the relative bias was nearly 400%. Shi et al. (2019) showed a similar result for the relative bias of the finite sample RMSEA at $n = 200$ and $K \approx 400$, but also demonstrated that the bias was practically negligible for $K \approx 25$. Furthermore, the RMSEA for a model with $K \approx 400$ achieved an acceptable relative bias around 10% at $n = 500$, implying that we could expect convergence for models smaller than this somewhere between these two sample sizes. As a result, the smallest model (CFA) had $K = 25$ degrees of freedom

and the largest (REPCFA) had $K = 133$. The LVPM model was chosen to fit between these two extremes at $K = 86$.

The models were represented, generally, in terms of the "all-y variant" of the LISREL parameterization (Jöreskog & Sörbom, 2006; Rosseel, 2012), given as follows:

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Theta})$$

$$\boldsymbol{\zeta} \sim N(\mathbf{0}, \boldsymbol{\Psi}),$$

where $\boldsymbol{\eta}$ are the latent variables. The vectorized parameters were represented collectively as $\boldsymbol{\theta} = [\mathrm{vec}(\boldsymbol{\Lambda})', \mathrm{vech}(\boldsymbol{\Theta})', \mathrm{vec}(\mathbf{B})', \mathrm{vech}(\boldsymbol{\Psi})']'$. This model implied the covariance structure

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B}')^{-1}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}.$$

For each model, a population covariance matrix $\boldsymbol{\Sigma}(\tilde{\boldsymbol{\theta}})$, with which the model had exact fit, was determined by the parameter values $\tilde{\boldsymbol{\theta}}$ defined below.

Model 1 (3-Factor CFA, $K = 25$)

The first model, based on a three-factor subset of the model from Holzinger and Swineford (1939), hypothesized that the covariance among nine observed variables $\mathbf{y} = [v_1, v_2, v_3, t_1, \ldots, s_3]'$ was the consequence of covariation among the three latent variables $\boldsymbol{\eta} = [V, T, S]'$ they measure. Each indicator loaded uniquely onto a single factor (e.g., v_i loads onto V with no cross-loadings), T covaried with V and S but they did not covary with each other, and none of the indicator error terms covaried. For clarity, Figure 1 displays the path diagram for the model. Given this specification, the model had $K = 25$ degrees of freedom.

The parameter values $\tilde{\boldsymbol{\theta}}$ defining the exact fit population were set as follows:

1) The latent variable covariance matrix was defined $\boldsymbol{\Psi} = \begin{bmatrix} 1 & & \\ 0.3 & 1 & \\ 0 & 0.3 & 1 \end{bmatrix}$.

2) Factor loadings were set to 0.75 (i.e., $\boldsymbol{\Lambda} = 0.75 * \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$).

3) The elements of the error term covariance matrix were defined $\theta_{ii} = 1 - 0.75^2$ and $\theta_{ij} = 0$, for all i and $j \neq i$.

I tested the proportion of misfit for the following three constraint subsets:

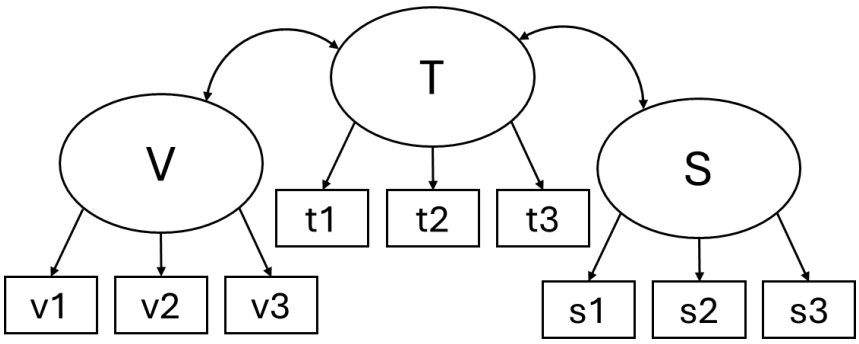
1) $S_{11} = \{\psi_{VS} = 0\}$ (omitted factor covariance) with $k = 1$ degrees of freedom.

2) $S_{12} = \{\lambda_{s_2V} = 0, \lambda_{t_2T} = 0, \lambda_{v_2S} = 0\}$ (omitted cross-loadings) with $k =$ 3 degrees of freedom.

3) $S_{13} = S_{11} \cup S_{12} \cup \{\theta_{v_1t_1} = 0, \theta_{t_3s_3} = 0\}$ (omitted error covariances) with $k =$ 6 degrees of freedom.

Figure 1

Path Diagram for Model 1 (3-Factor CFA, K=25)



Note. Error terms omitted for clarity. *K* = model degrees of freedom.

Model 2 (5-Factor LVPM, $K = 86$)

The second model hypothesized that fifteen indicators $\mathbf{y} = [a_1, a_2, a_3, b_1, \ldots, e_3]'$ each measured one of five latent variables $\boldsymbol{\eta} = [A, B, C, D, E]'$. Furthermore, it specified the set of linear structural relations among the $\boldsymbol{\eta}$ depicted in Figure 2. Each indicator measured a single factor (e.g., a_i loaded onto A with no cross-loadings), the exogenous factors (A and D) did not covary, and none of the latent or observed variable error terms covaried. In total, the model had $K = 86$ degrees of freedom.

The parameter values $\widetilde{\boldsymbol{\theta}}$ defining the exact fit population were set as follows:

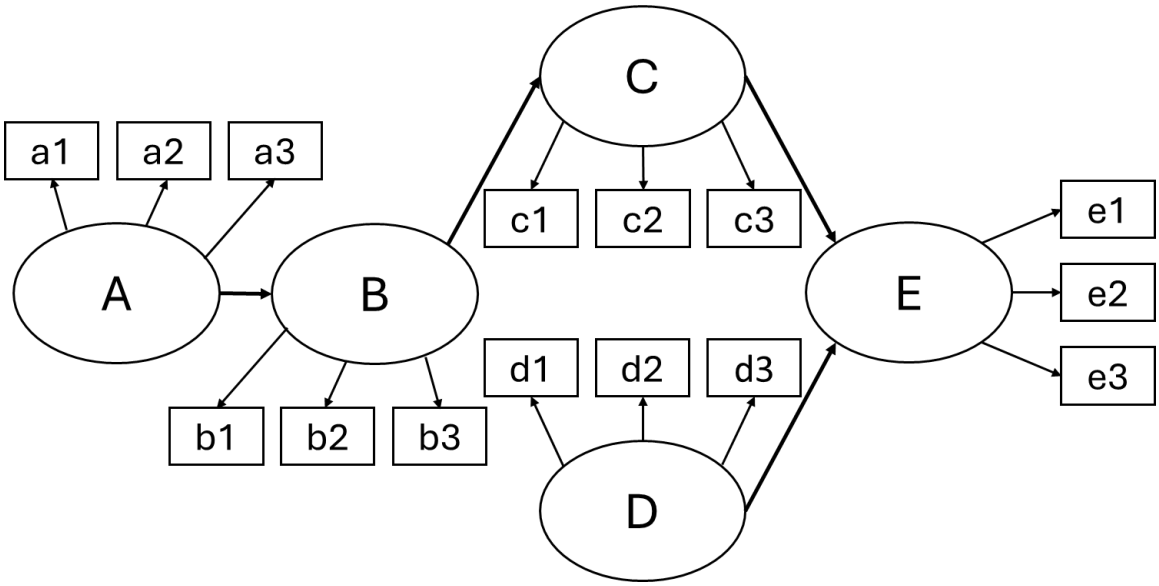
1) The path coefficients were set to $B_{BA} = B_{CB} = 0.5$, $B_{EC} = B_{ED} = 0.35$, and zero otherwise
2) The latent variable covariance matrix was defined with $\psi_{AA} = \psi_{DD} = 1$, $\psi_{BB} = \psi_{CC} = 1 - 0.5^2$, and $\psi_{EE} = 1 - 2 * 0.35^2$.
3) Factor loadings were set to 0.75 and cross-loadings equal to 0.
4) The elements of the error term covariance matrix were defined $\theta_{ii} = 1 - 0.75^2$ and $\theta_{ij} = 0$, for all i and $j \neq i$.

I tested the proportion of misfit for the following three constraint subsets:

1) $S_{21} = \{B_{DB} = 0\}$ (omitted path coefficient) with $k = 1$ degrees of freedom.
2) $S_{22} = \{\psi_{ij} = 0 \mid i \neq j\}$ (omitted exogenous factor and factor error term covariances) with $k = 6$ degrees of freedom.
3) $S_{23} = S_{21} \cup S_{22} \cup \{\theta_{a_i e_i} = 0, \theta_{b_i e_i} = 0, \theta_{c_i d_i} = 0 \mid \forall\, i\}$ (omitted error covariances for the indicators of A and E, B and E, and C and D that had the same index) with $k = 15$ degrees of freedom. (Note that this subset is overparameterized. Given the redundancy of B_{DB} and ψ_{DB}, it contains one more constraint than degrees of freedom)

Figure 2

Path Diagram for Model 2 (5-Factor LVPM, $\boldsymbol{K = 86}$)



Note. Error terms omitted for clarity. *K* = model degrees of freedom.

Model 3 (6-Factor REPCFA, $K = 133$)

The third model extended the first to a repeated measures CFA, with an additional three-factor structure representing measurement at a second time point. Indicators $\mathbf{y}$ and factors $\boldsymbol{\eta}$ were differentiated at each time point by their index – 1 for the first measure, 2 for the second (e.g., V_1 is from time 1 and s_{23} is from time 2). The model hypothesized the same measurement and structural components at each time point. Furthermore, it specified that factors in the first submodel covaried with their counterparts in the second. Finally, the model constrained all between-time error term covariances and cross-loadings to 0, loadings at time 1 to be equal to their counterparts at time 2 (e.g., $\lambda_{v_{11}} = \lambda_{v_{21}}$), and nonzero within-time covariances at time 1 to be equal to their counterparts at time 2 (e.g., $\psi_{V_1T_1} = \psi_{V_2T_2}$, but not $\psi_{V_1S_1} = \psi_{V_2S_2}$). The path diagram in Figure 3 provides additional clarity for the specifics of the model specification. This model had $K = 133$ degrees of freedom.

In addition to the specifications described above for the first model, the parameter values $\widetilde{\boldsymbol{\theta}}$ defining the exact fit population are given as follows:

1) The latent variable covariance matrix was defined such that factors at time 1 covaried with their counterpart at time 2: $\psi_{V_1V_2} = \psi_{T_1T_2} = \psi_{S_1S_2} = 0.8$.

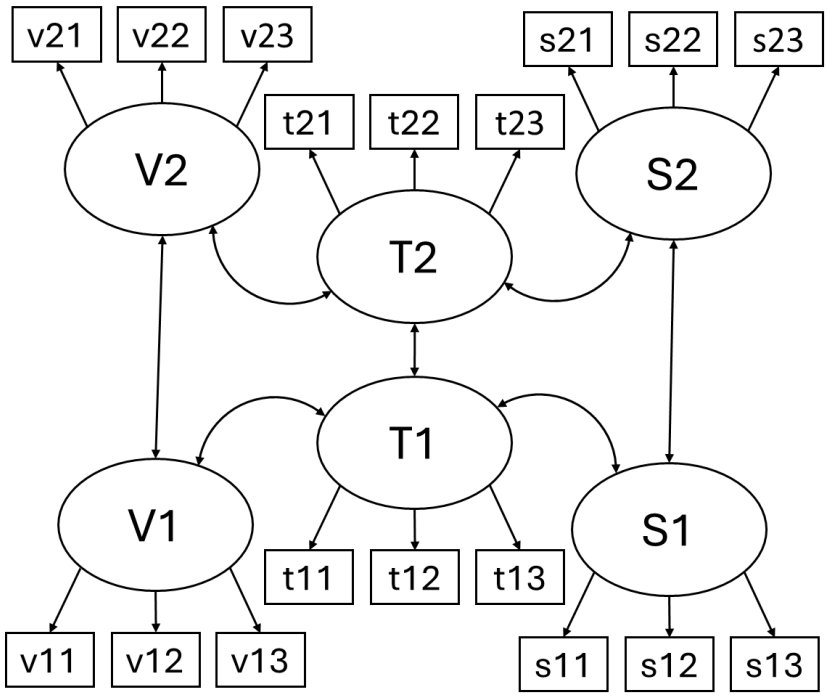
I tested the proportion of misfit for the following three constraint subsets:

1) $S_{31} = \{\psi_{V_2S_2} = 0\}$ (omitted factor covariance at time 2) with $k = 1$ degrees of freedom.
2) $S_{32} = \{\lambda_{V_1v_{1i}} = \lambda_{V_2v_{2i}}, \lambda_{T_1t_{1i}} = \lambda_{T_2t_{2i}}, \lambda_{S_1s_{1i}} = \lambda_{S_2s_{2i}} \mid \forall i\}$ (loading equality constraints) with $k = 9$ degrees of freedom.

3) $S_{33} = S_{31} \cup S_{32} \cup \{\psi_{V_1T_1} = \psi_{V_2T_2}, \psi_{T_1S_1} = \psi_{T_2S_2}\} \cup \{\theta_{v_{1i}v_{2i}} = 0, \theta_{t_{1i}t_{2i}} = 0, \theta_{s_{1i}s_{2i}} = 0 \mid \forall\, i\}$ (time invariance constraints on factor covariances and omitted between-time error covariances) with $k = 21$ degrees of freedom.

Figure 3

Path Diagram for Model 3 (6-Factor REPCFA, $\boldsymbol{K = 133}$)



Note. Error terms omitted for clarity. *K* = model degrees of freedom.

Sampling and Estimation of Misfit Proportion Statistics

Simulated datasets were generated and analyzed using custom code and freely-available software packages in the *R* programming language (R Core Team, 2024). Data for each simulation study condition were generated separately for each model and constraint subset. The populations from which the data were sampled had the same implied covariance matrix $\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}})$ for every replication in the exact fit conditions, and $\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}) + \mathbf{E}$ for the conditions with approximation error, where $\mathbf{E}$ is generated randomly for each replication as outlined below. Then, a sample was drawn from a multivariate normal distribution with the population covariance matrix using the mvtnorm package (Genz & Bretz, 2009). I fit the base model and the modified model with normal theory maximum likelihood estimator implemented in the `lavaan` package (Rosseel, 2012). The misfit proportion statistic T_{diff}/T was then computed from fit statistics for both estimates. Outcome measures for both studies were computed and analyzed using base *R* functionality and packages from the tidyverse collection (Wickham et al., 2019).

Generating Populations with Approximation Error

For conditions in which the model was meant to be misspecified, I defined the population covariance matrix as a perturbation of the exact fit covariance matrix for the model – that is, $\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}) + \mathbf{E}$. The perturbation matrix $\mathbf{E}$ was generated randomly following a procedure adapted from Lai (2019). Given a target amount of discrepancy F_0 the process for sampling $\mathbf{E}$ described in this paper satisfied the following conditions:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, F\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}) + \mathbf{E}\right) = \widetilde{\boldsymbol{\theta}}$$

$$F\left(\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}), \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}) + \mathbf{E}\right) = F_0.$$

In plain terms, these conditions imply that, if the model were fit to the population covariance matrix associated with an $\mathbf{E}$ generated by this method, the parameter estimate would be $\widetilde{\boldsymbol{\theta}}$ with global discrepancy F_0.

I adapted the method to control the proportion of approximation error π_{diff} associated with a targeted constraint subset. Modifying the model by releasing this constraint subset and refitting it to the population covariance matrix imposed a third restriction on $\mathbf{E}$:

$$\min_{\boldsymbol{\theta}} F\left(\boldsymbol{\Sigma}_*(\boldsymbol{\theta}), \boldsymbol{\Sigma}\left(\widetilde{\boldsymbol{\theta}}\right) + \mathbf{E}\right) = F_0(1 - \pi_{\text{diff}}),$$

where $\boldsymbol{\Sigma}_*(\boldsymbol{\theta})$ was the estimated covariance structure from the modified model. Additionally, to emulate even dispersion of misfit, the adapted method sampled $\mathbf{E}$ approximately uniformly with respect to F and $\boldsymbol{\Sigma}\left(\widetilde{\boldsymbol{\theta}}\right)$ from all matrices that satisfied these conditions. Despite the restrictions on $\mathbf{E}$, there are infinitely many solutions, determined primarily by the initial value chosen for the minimization procedure. A comprehensive description of the method for generating initial values for $\mathbf{E}$ that result in approximate elliptical uniformity can be found in Appendix A. Values for F_0 and π_{diff} were determined by the simulation study conditions defined in the next section.

Study Design

Study 1: Empirical False Positive Rate

The first simulation study evaluated the empirical false positive rate of the misfit proportion test for all nine model and modification pairs across the following manipulated population and sample conditions:

1) The size of the samples drawn from each population were one of $n = 101, 201, 401, 801,$ or 1601.

2) Populations were generated for the model to have exact fit or evenly dispersed approximation error $F_0 = \epsilon^2 K$, where $RMSEA\ \epsilon = 0.00, 0.03, 0.05, 0.08,$ or 0.10.

To ensure the proportion of approximation error associated with the constraint subset was consistent with evenly dispersed misfit, π_{diff} was sampled from $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ for each replication.

I selected the values for these manipulated factors to cover typical sample sizes and RMSEA benchmarks. Prior studies of the model size effect on the LR test statistic (Moshagen et al., 2012; Shi et al., 2019), discussed earlier, found finite sample bias for the null distribution vanishes around $N = 200 - 500$ observations. However, approximation error also contributes to this bias (Yuan et al., 2007), implying that larger samples may be needed for convergence under approximate fit conditions ($RMSEA > 0.00$). Consequently, I settled on the sample size range $N = 101 - 1601$, aligning with $N = 100 - 1000$ from Shi et al (2019) and $N = 150 - 5000$ from Hu & Bentler (1999).

The RMSEA values correspond to the foundational benchmarks proposed by Cudeck & Brown (1993): 0.00 (exact fit), 0.05 (close fit), and 0.08 (reasonable fit). The maximum value, 0.10, generally represents the threshold for *poor fit* (MacCallum, 1996; Jackson, 2009). The level of 0.03 was incorporated as an intermediate value between exact and close fit.

Including the number of models and modifications, the total number of manipulated conditions in the fully factorial design was $25 \times 9 = 225$. In each cell, empirical false positive rate $\hat{\alpha}$ was computed for three asymptotic false positive rates $\alpha = 0.10, 0.05,$ and 0.01. The estimate of $\hat{\alpha}$ was defined as

$$\hat{\alpha} = \frac{1}{R}\sum_{i=1}^{R} I\{(T_{\text{diff}}/T)_i \geq Q_\alpha\},$$

where Q_α is the $(1-\alpha)$ quantile of the $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ distribution.

To control the Monte Carlo error, I determined the number of replications R by modeling the null distribution of the $\hat{\alpha}$ given α. Assuming the misfit proportion statistics had converged to their asymptotic null distribution, it follows that $R\hat{\alpha} \sim \text{Binomial}(R, \alpha)$. Because the precision of $\hat{\alpha}$ under this model depended on α, I defined R based on the relative bias implied by the 95% confidence interval (CI) of $\hat{\alpha}$ at $\alpha = 0.01$, the value with the least precise estimate. I chose $R = 40000$ such that the CI would be approximately (0.009, 0.011), corresponding to $\pm$10% relative bias. This number of replications produced tighter intervals for higher α levels: (0.048, 0.052) for $\alpha = 0.05$ ($\pm$4% relative bias) and (0.097, 0.103) at $\alpha = 0.10$ ($\pm$3% relative bias).

Study 2: Empirical Power

The second study evaluated the empirical power of the misfit proportion test. Just as with the first, this study manipulated sample size and population approximation error. However, in this case the proportion of misfit π_{diff} was fixed for each replication to a value determined by the sample size n, the population approximation error F_0, and the target nominal power $(1-\beta)$. The levels that these factors took were as follows

1) The size of the samples drawn from each population were one of $n = 101, 201, 401, 801,$ or 1601.
2) The population approximation error were set to values $F_0 = \epsilon^2 K$, where $RMSEA\ \epsilon = 0.03, 0.05, 0.08,$ or 0.10.
3) Given n and ϵ, the effect size π_{diff} was set to the value that attained asymptotic power of $(1-\beta) = 0.5$ or 0.8 as described below.

Given $\delta = (n-1)F_0$ and π_{diff}, the asymptotic power for the misfit proportion test of a constraint subset with k degrees of freedom from a model with K degrees of freedom is given by

$$(1-\beta) = 1 - F(Q_\alpha),$$

where Q_α is the α quantile for the central beta distribution and F is the distribution function for $Beta\left(\frac{k}{2}, \frac{K-k}{2}; \delta\pi_{\text{diff}}, \delta(1-\pi_{\text{diff}})\right)$, the doubly noncentral beta distribution (Orsi, 2017). Thus, for each cell, π_{diff} was obtained by finding the root of $F(Q_\alpha) - \beta = 0$ using the uniroot function in the base distribution of the *R* programming language (R Core Team, 2024). The misfit proportion statistic could not achieve the targeted asymptotic power level in some conditions. As such, I dedicated a subsection to review the results of the asymptotic power analysis and provided plots illustrating how the different models, constraint subsets, and simulation conditions impacted π_{diff}.

In the fully factorial design, the total number of manipulated conditions was $40 \times 9 = 360$. Empirical power was computed at only a single asymptotic false positive rate, $\alpha = 0.05$. Its estimate was defined as

$$\left(1-\hat{\beta}\right) = \frac{1}{R}\sum_{i=1}^{R} I\{(T_{\text{diff}}/T)_i \geq Q_\alpha\}$$

where Q_α is the 0.95 quantile of the $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ distribution.

The number of replications R was determined using the same approach as in Study 1. First, the null distribution of the Monte Carlo error for $\hat{\beta}$ given the nominal power level $(1-\beta)$ was taken to be $R\hat{\beta} \sim \text{Binomial}(R, \beta)$. For the smallest $\beta = 0.2$, the 95% CI corresponding to $\pm$10% relative bias occurred at $R = 1550$. However, I chose $R = 10000$ because it was relatively fast to generate the replications, and I preferred the higher precision. At this number of

replications, the 95% CI would be (0.19, 0.21) at $\beta = 0.2$ ($\pm$5% relative bias) and (0.49, 0.51) at $\beta = 0.5$ ($\pm$2% relative bias).

Analytical Methods

Regression Analysis

To probe the impact of the manipulated factors on the convergence of the misfit proportion statistics, I fit a regression model to the finite sample biases: $\hat{\alpha} - \alpha$ in Study 1 and $\beta - \hat{\beta}$ in Study 2. Because the effects were uniform across the nominal α and β levels, I limited my focus to $\alpha = 0.05$ and $\beta = 0.2$ (0.8 nominal power). The model estimated two independent components of convergence: one that varied with the idiosyncrasies of the model forms and modification subsets, and one that depended on the amount of approximation error. The first component was modeled with a three-way interaction between model form, modification subset, and sample size. The second was modeled with a two-way interaction between approximation error and sample size. All main effects and lower-order interactions were included for completeness.

Because I expected convergence to be nonlinear in sample size, I treated it as a categorical variable, estimating separate intercepts for each level. I modeled approximation error as a linear effect to constrain the estimation at low levels of approximation error, as the statistics could not achieve one or both target power levels in some conditions in Study 2. I extended this specification to the Study 1 model for the sake of parity and comparison.

This approach to specification was admittedly *ad hoc*. My goal was not to draw statistically rigorous inferences about the convergence of these statistics, but rather merely to provide some structure to the eventual interpretation of the results. As such, I reported the size of

significant effects in terms of partial η^2 with the thresholds for small, medium, and large effects being 0.01, 0.09, and 0.26 (Cohen et al., 2003), respectively. To clarify the meaning of the effects, I tested pairwise mean and trend comparisons using the `emmeans` package in R (Lenth, 2025). To account for multiplicity, *p*-values were adjusted using Scheffé's method. For the significant comparisons, I reported the means (M) or slopes and standard errors (SE) for each group and the *p*-value for the test. Because only the modification subsets with $k = 1$ degrees of freedom were comparable between models, I did not report comparisons between different modification subsets from different model form conditions.

Convergent Sample Size Analysis

Convergence of T_{diff}/T within each cell was evaluated the using a $\pm$10% relative bias threshold for practical significance. This criterion is substantially more stringent than the $\pm$50% liberal threshold recommended by Bradley (1978) for studies like this and aligns with the $\pm$10% threshold employed by Shi et al. (2019) to detect significant model size effects for RMSEA. Estimates that fell outside this range indicated that the simulated sample meaningfully diverged from the corresponding asymptotic distribution.

For Study 1, this yielded practical significance intervals that matched the 95% CI of $\hat{\alpha}$ at $\alpha = 0.01$, (0.009, 0.011) but were wider for the other nominal significance levels – namely, (0.045, 0.055) for $\alpha = 0.05$ and (0.09, 0.11) for $\alpha = 0.10$. Ignoring dependence among the $\hat{\alpha}$ at the multiple α levels, the largest false positive rate for practical significance was at most 5% under the binomial model of Monte Carlo error, given $R = 40000$.

In study 2, the practical significance intervals for $(1-\hat{\beta})$ were (0.78, 0.82) at $(1-\beta)=0.8$ and (0.45, 0.55) at $(1-\beta)=0.5$. Under the binomial model, the false positive rate for practical significance was less than 0.001% at $R=10000$ for either nominal power level.

The goal of this analysis was to identify a sample size threshold for the misfit proportion statistics, beyond which it could be used in practice without concern for the impact of finite sample errors. For each combination of model, modification subset, and approximation error conditions, this coincided with the point at which the size and number of practically significant biases became negligible.

Results

Study 1: Empirical False Positive Rate

Regression Analysis. The results for Study 1 are shown in Figures 4, 5, and 6 for the CFA, LVPM, and REPCFA conditions, respectively. Tables with the raw data behind these figures can be found in Appendix B. The effects of the manipulated simulation conditions on the absolute relative bias of the estimated empirical alphas were predominantly uniform across the nominal significance levels, so I will focus on the results for $\alpha=0.05$. In Table 1, I report the results of the regression.

The analysis supported the hypothesis that the model size effect in the LR statistics in both the numerator and denominator contributed to the finite sample bias in the misfit proportion test. A large main effect of model form ($\eta^2=0.57$) was observed, indicating that the bias decreased as a function of the model degrees of freedom K. Specifically, the average relative bias in the CFA model condition ($K=25$; $M=2.99\%$, $SE=0.30\%$) was significantly more positive ($p<0.001$) than in the larger REPCFA model condition ($K=133$; $M=$

-2.73%, $SE = 0.30\%$). The bias in the LVPM model ($K = 86$; $M = 2.66\%$, $SE = 0.30\%$) form condition did not significantly differ from the other two models. Furthermore, the large interaction effect with sample size ($\eta^2 = 0.61$) indicated that the model size effect was most pronounced at low sample sizes. At $n = 101$, the average relative bias in the CFA condition ($M = 7.63\%$, $SE = 0.67\%$) and LVPM condition ($M = 4.30\%$, $SE = 0.67\%$) were both significantly more positive ($p < 0.001$) than in the REPCFA condition ($M = -10.9\%$, $SE = 0.67\%$). The bias in the CFA and LVPM conditions did not significantly differ. By $n = 1601$, the average relative bias was near zero in the CFA ($M = 0.78\%$, $SE = 0.67\%$), LVPM ($M = 1.74\%$, $SE = 0.67\%$), and the REPCFA conditions ($M = 0.60\%$, $SE = 0.67\%$), suggesting the model size effect began to vanish at higher sample sizes as the test statistics converged to their asymptotic distributions.

As expected, the model size effect specifically coming from the numerator statistic had a noticeably smaller impact. There was a small main effect of modification subset ($\eta^2 = 0.05$). Averaged across model form and sample size conditions, the relative bias was significantly more positive ($p = 0.020$) in modification subset 1 ($k = 1$ for all models; $M = 1.22\%, 0.30\%$) and modification subset 3 ($k = 6, 15, 21$, respectively; $M = 0.24\%, 0.30\%$). Because the modification subsets differ in size between model form conditions, the moderate interaction effect ($\eta^2 = 0.16$) was a more accurate measure of the model size effect from the numerator. In the LVPM model form condition, the average relative bias in modification subset 2 ($k = 6;\ M = 4.49\%$, $SE = 0.52\%$) was significantly larger ($p = 0.024$) than modification subset 1 ($k = 1;\ M = 1.36\%$, $SE = 0.52\%$). In contrast, the REPCFA model form condition, the average relative bias in modification subset 1 ($k = 1;\ M = -0.97\%$, $SE = 0.52\%$) was significantly more positive ($p = 0.011$) than modification subset 3 ($k = 21;\ M = -4.31\%$, $SE = 0.52\%$).

Focusing on modification subsets with $k = 1$ degrees of freedom, the bias in the CFA model form condition ($M = 3.26\%$, $SE = 0.52\%$) was significantly more positive ($p < 0.001$) than for the REPCFA model ($M = -0.97\%$, $SE = 0.52\%$), but neither was significantly different than the bias in the LVPM model ($M = 1.36\%$, $SE = 0.52\%$). The moderately-sized three-way interaction effect ($\eta^2 = 0.15$) highlighted how much larger this difference was at lower sample sizes. At $n = 101$, the average relative bias in modification subset 1 was significantly more positive ($p = 0.039$) in the CFA model form condition ($M = 6.67\%$, $SE = 1.16\%$) than the REPCFA condition ($M = -6.57\%$, $SE = 1.16\%$), although neither was significantly different than the LVPM model ($M = 2.01\%$, $SE = 1.16\%$). As with the lower-order model form and sample size interaction effect, the average relative bias for modification subset 1 became negligible by $n = 1601$ in the CFA ($M = 1.03\%$, $SE = 1.16\%$), LVPM ($M = 0.88\%$, $SE = 1.16\%$), and the REPCFA conditions ($M = 1.44\%$, $SE = 1.16\%$). Furthermore, the differences between modification subsets did not vary significantly by sample size.

Finally, the approximation error effect was large ($\eta^2 = 0.48$), with the bias significantly increasing linearly by 3.22% ($SE = 0.49\%, p < 0.001$) from $RMSEA = 0.00$ to 0.10. The medium interaction with sample size ($\eta^2 = 0.23$) indicated that this biasing effect generally significantly decreased ($p < 0.001$) with sample size. At $n = 101$, the change in bias was 12.4% ($SE = 1.09\%, p < 0.001$), whereas the change at $n = 1601$ was 2.6% ($SE = 1.09\%, p = 0.328$).

Convergent Sample Size Analysis. Across all conditions, the statistics appeared to have not converged at a sample size of $n = 101$, but most converged by $n = 201$. There were only a few conditions in which the bias exceeded the practical significance criterion of 10% at greater sample sizes, but the vast majority occurred at the smallest samples sizes. The size and

directionality of the practically significant biases depended on many factors, but most notably by model form.

In the CFA model form condition ($K = 25$), the practically significant biases were positive, indicating that the false positive rate was higher than expected in those conditions. Despite exceeding the conservative practical significance threshold, the largest relative bias was small, less than 15% of the nominal significance level. These biases occurred primarily in the conditions with the smallest sample size of $n = 101$ and the two largest approximation error conditions ($\epsilon = 0.08$ and 0.10). The tests of the smallest ($k = 1$) and largest ($k = 6$) modification subsets had most of the significant biases. Additionally, there were a few isolated biases at larger sample sizes. In the $\epsilon = 0.08$ condition, the smallest modification subset also had significant biases for $n = 401$ and 801. Additionally, there was one significant bias in the exact fit condition ($\epsilon = 0.00$) at $n = 101$ for the largest modification subset.

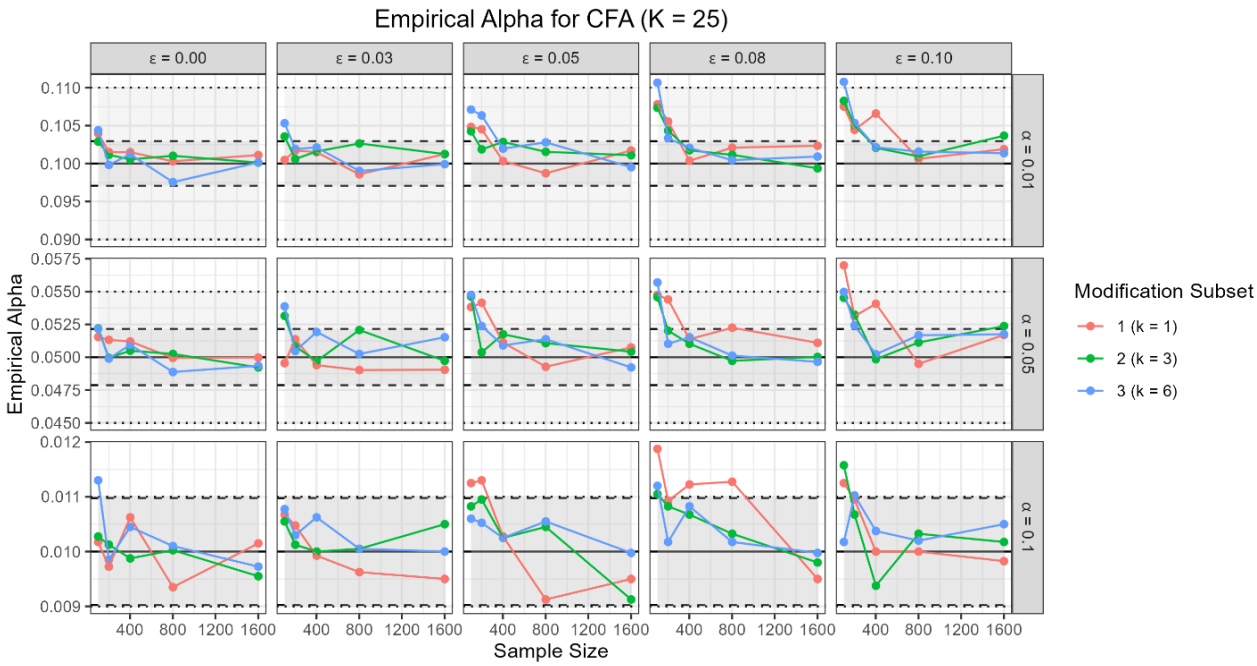
Similarly, the practically significant biases in the LVPM model form condition ($K = 86$) were positive. The largest relative biases approached 30%. Again, these biases occurred in the conditions with the smallest sample size and largest approximation error. Most arose from the tests of the second modification subset ($k = 6$). There were a few isolated significant biases, including at some middling sample sizes ($n = 201$ and 401) in the largest approximation error condition ($\epsilon = 0.10$), and a couple in the tests of the other two modification subsets ($k = 10$ and 15) at $n = 801$.

In contrast, the REPCFA model form condition ($K = 133$) had practically significant biases that were negative, indicating false positive rates that were lower than expected. Furthermore, these biases were the largest at the smallest sample size ($n = 101$) in the exact fit condition ($\epsilon = 0.00$) and trended smaller as the approximation error increased. Notably, the

magnitude of these biases increased with the size of the modification subset. The largest biases were comparable in size to those from the LVPM model form condition, with the maximum being nearly 33%. A few of the practically significant biases were positive, occurring in the tests of the first modification subset ($k = 1$) at middling sample sizes ($n = 201$ and 801). Similarly, there were one or two small significant biases at $n = 801$ for the other two modification subset tests ($k = 9$ and 21). However, these isolated biases at larger sample sizes were quite small, barely breaking 10%.

Figure 4

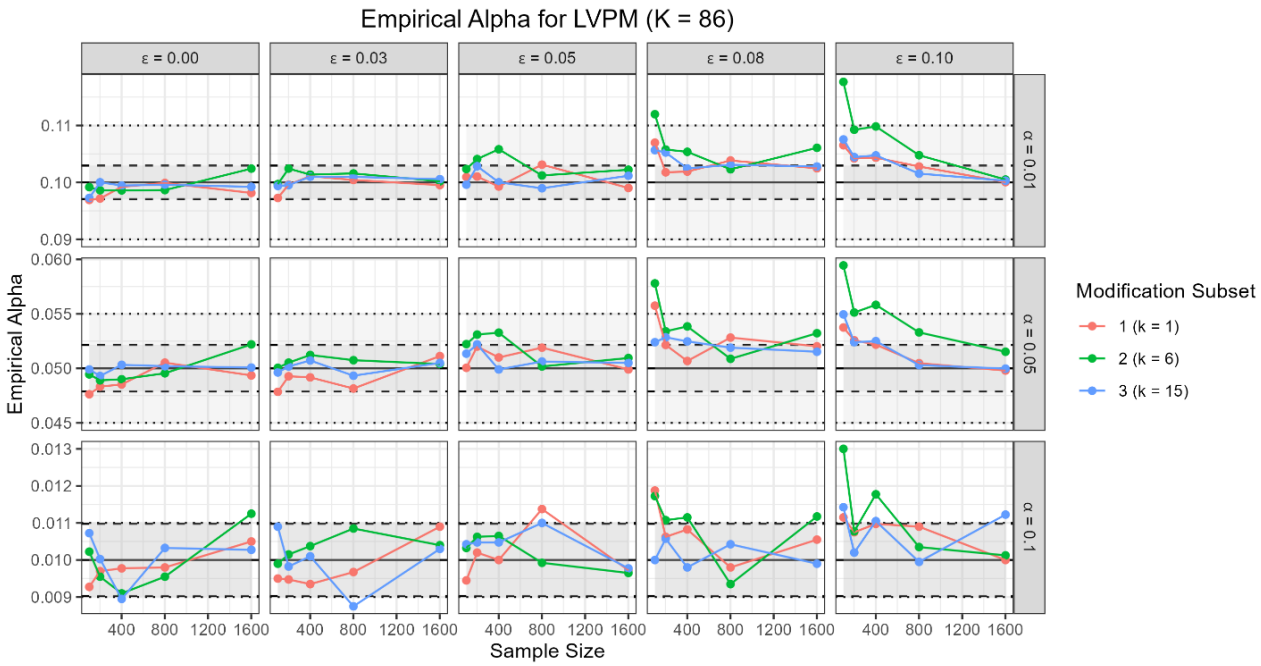
Empirical false positive rate (alpha) for Model 1 (CFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), α = theoretical false positive rate.

Figure 5

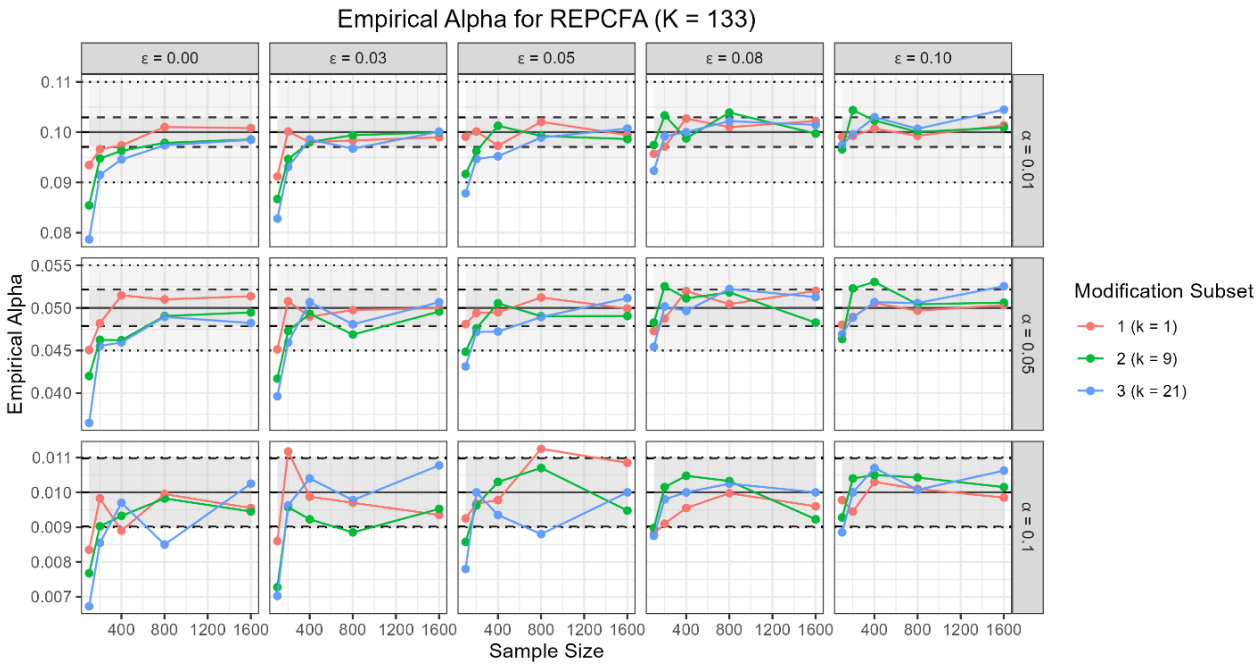
Empirical false positive rate (alpha) for Model 2 (LVPM)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), α = theoretical false positive rate.

Figure 6

Empirical false positive rate (alpha) for Model 3 (REPCFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), α = theoretical false positive rate.

Table 1

Effects of manipulated factors on finite sample bias in empirical false positive rate

Source	df	*F*	*p*	partial η^2
Model Form (Form)	2	114.95	< 0.001	0.57††
Modification Subset (Subset)	2	4.54	0.012	0.05
Sample Size (n)	4	1.38	0.242	0.03
Form × Subset	4	8.10	< 0.001	0.16†
Form × n	8	34.55	< 0.001	0.61††
Subset × n	8	0.80	0.604	0.04
Form × Subset × n	16	1.97	0.017	0.15†
Approximation Error (ϵ)	1	162.95	< 0.001	0.48††
$\epsilon \times n$	4	13.17	< 0.001	0.23†

Note. All factors are modeled as categorical variables, with separate indicators for each non-reference level, except Approximation Error (ϵ), which was modeled as a linear relation. Daggers denote the Cohen et al. (2003) partial η^2 thresholds for multiple regression delineating small-medium and medium-large sized effects, respectively,

†$0.09 \leq \eta^2 < 0.26$, ††$0.26 \leq \eta^2$

Study 2: Asymptotic Power Analysis

Figures 7, 8, and 9 show the values for the π_{diff} parameter that achieved the target asymptotic power for the CFA, LVPM, and REPCFA conditions, respectively. Tables with the raw data behind these figures can be found in Appendix C. The π_{diff} values decreased as a function of the noncentrality parameter of the global model fit statistic, which depended on both sample size and global RMSEA. It appeared to converge toward the Q_α, the 0.95 quantile of the $Beta\left(\frac{k}{2}, \frac{K-k}{2}\right)$ distribution, where K is the total model degrees of freedom and k is the degrees of freedom for the modification subset. Additionally, the π_{diff} value scaled with the proportion of the modification subset degrees of freedom relative to the model degrees of freedom, increasing with k/K. Furthermore, a larger π_{diff} was required to achieve a power of 0.8 than 0.5.

In several conditions, there did not exist a value for π_{diff} that could achieve one or both target power levels. This occurred in conditions with small sample size and small approximation error, in which we would expect the evenly dispersed sampling variability to overwhelm the unevenness of the approximation error. Because of this, tests computed in such conditions would be underpowered to detect deviations from evenly dispersed misfit, even if all the approximation error were concentrated in the selected modification subset (i.e., $\pi_{\text{diff}} = 1$). These cells were omitted from the empirical power study.

To briefly review these results, let us limit our focus to those with a nominal power level of 0.8. For the CFA model form condition ($K = 25$), the test of the smallest modification subset ($k = 11$) was only underpowered in the condition where $\epsilon = 0.03$ and $n = 101$. The tests of the two largest modification subsets ($k = 3$ and 6) were underpowered when approximation error ($\epsilon = 0.03$ and 0.05) at the and sample size ($n = 101$ and 201) were small. Furthermore, the

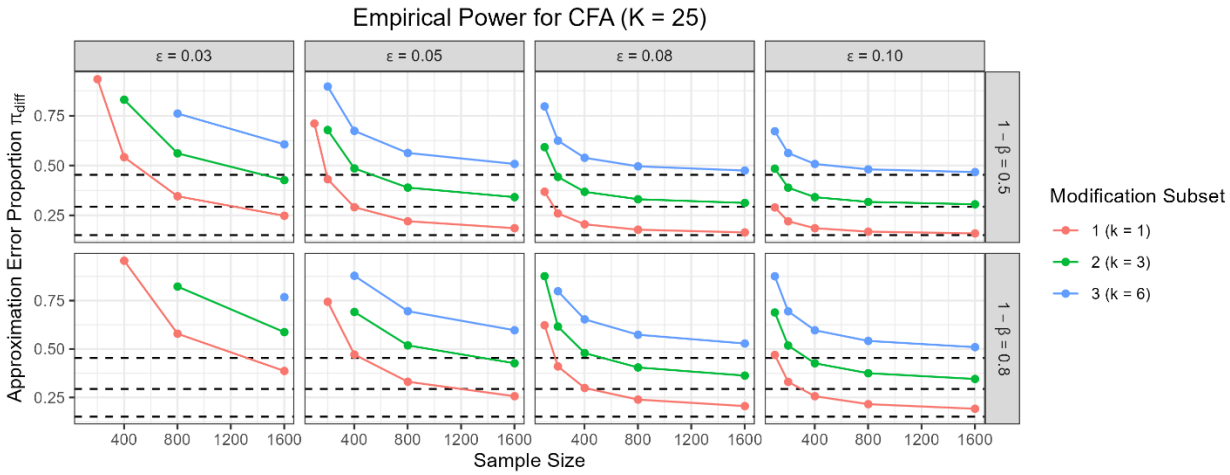
largest modification subset was underpowered at $n = 401$ and 801 when $\epsilon = 0.03$, although it was quite close ($1 - \beta = 0.79$) in the $n = 801$ condition.

The LVPM model form condition ($K = 86$) had fewer instances in which its tests were underpowered. The tests for both of the smallest modification subsets ($k = 1$ and 6) were able to achieve a nominal power of 0.8 in all conditions. The largest modification subset ($k = 9$) was only underpowered at $n = 101$ and $\epsilon = 0.03$.

In the REPCFA model form condition ($K = 133$), the only underpowered tests were in the $\epsilon = 0.03$ condition. Both tests of the modification subsets 2 and 3 ($k = 9$ and 15) were underpowered at $n = 101$, and modification subset 3 was underpowered at $n = 201$ as well.

Figure 7

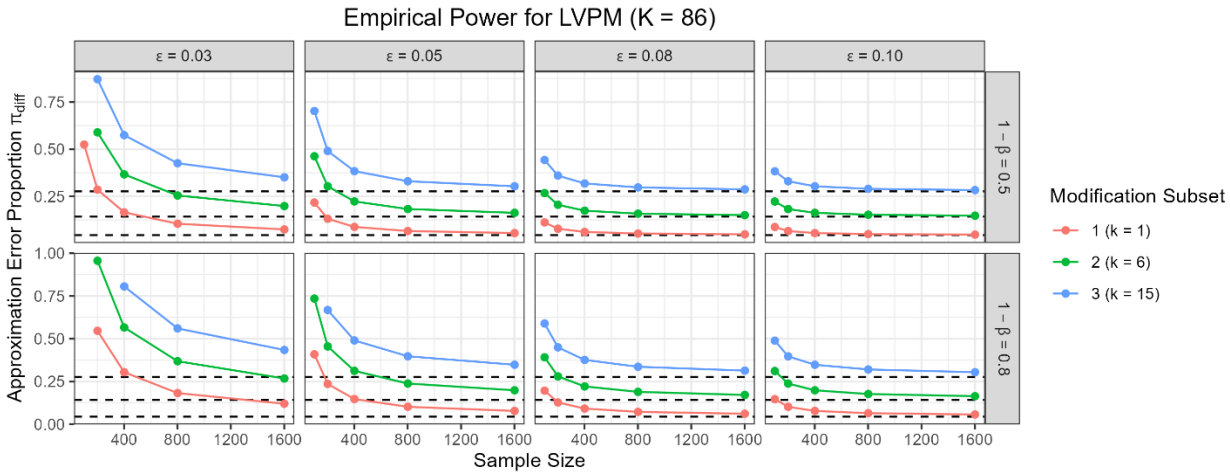
Approximation error proportion (π_{diff}) achieving target power for Model 1 (CFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Figure 8

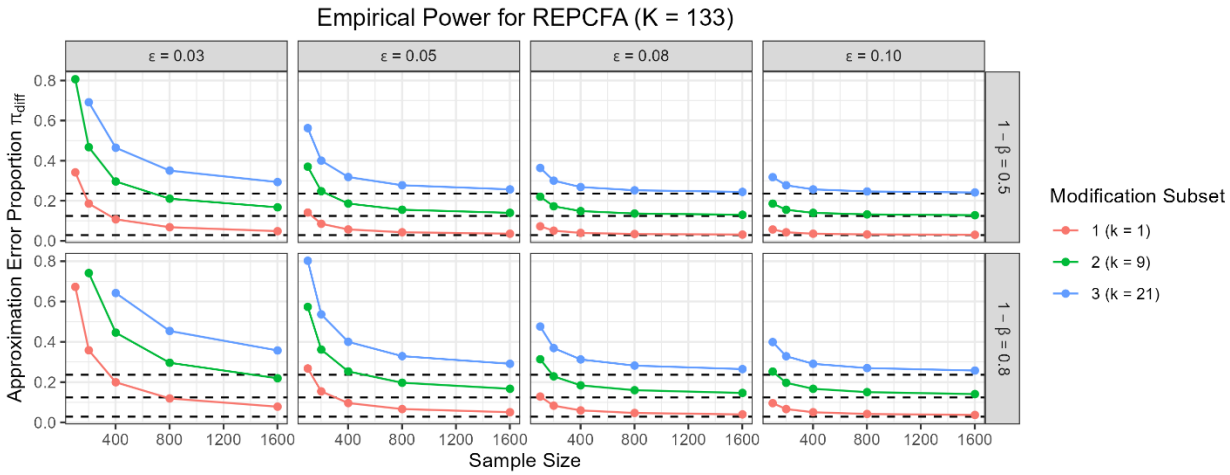
Approximation error proportion (π_{diff}) achieving target power for Model 2 (LVPM)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Figure 9

Approximation error proportion (π_{diff}) achieving target power for Model 3 (REPCFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Study 2: Empirical Power

Regression Analysis. Figures 10, 11, and 12 display the results of Study 2 for the CFA, LVPM, and REPCFA conditions, respectively. The corresponding data are reported in Appendix D. As with Study 1, the convergence of the misfit proportion statistics was similar between the two nominal power levels. As such, I will limit our focus to the report the results for the nominal 0.80 power level ($\beta = 0.20$). The results for the regression are presented in Table 2.

These results indicated that the model size effect also impacted the empirical alternative distribution of the misfit proportion statistics, though the size was much smaller than for the null distribution. The main effect of model form was moderate in size ($\eta^2 = 0.12$), such that the average relative bias in the LVPM model condition ($K = 86$; $M = -3.60\%$, $SE = 0.35\%$) was significantly more positive ($p = 0.011$) than in the larger REPCFA model condition ($K = 133$; $M = -5.07\%$, $SE = 0.33\%$). The bias in the CFA model ($K = 25$; $M = -4.56\%$, $SE = 0.43\%$) form condition did not significantly differ from the other two models. The interaction effect with sample size was also large ($\eta^2 = 0.20$); however, the differences between models did not significantly vary by sample size. At $n = 101$, the average relative bias was substantially negative in the CFA condition ($M = -11.0\%$, $SE = 1.34\%$), LVPM ($M = -13.0\%$, $SE = 0.95\%$), and REPCFA ($M = -17.0\%$, $SE = 0.80\%$). By $n = 1601$, the average relative bias was near zero in the CFA ($M = -0.52\%$, $SE = 0.72\%$), LVPM ($M = 0.23\%$, $SE = 0.72\%$), and the REPCFA conditions ($M = 0.52\%$, $SE = 0.72\%$).

The model size effect on the numerator statistic was even smaller, still, with the main effect of modification subset being insignificant. However, the interaction between model form and modification subset was moderate in size ($\eta^2 = 0.16$). In the REPCFA model form condition, the average relative bias in modification subset 1 ($k = 1$; $M = -3.23\%$, $SE =$

0.56%) was significantly more positive ($p = 0.010$) than modification subset 3 ($k = 21;\ M = -7.01\%, SE = 0.60\%$). Unlike Study 1, the bias in modification subsets with $k = 1$ degrees of freedom did not significantly vary by model form. The three-way interaction effect was moderate ($\eta^2 = 0.20$); however, none of the mean comparisons were that significant.

Finally, the approximation error effect was moderate ($\eta^2 = 0.13$), such that the bias significantly increasing linearly by 4.39% ($SE = 0.87\%, p < 0.001$) from $RMSEA = 0.00$ to 0.10. There was also a moderate interaction with sample size ($\eta^2 = 0.13$) suggesting that this biasing effect generally significantly decreased ($p = 0.018$) with sample size. At $n = 101$, the change in bias was 12.3% ($SE = 2.69\%, p < 0.001$), whereas the change at $n = 1601$ was 1.3% ($SE = 1.54\%, p = 0.406$).

Convergent Sample Size Analysis. In most conditions, the statistics appear to have converged by $n = 201$. No biases exceeded 10% absolute relative bias beyond this point, and very few biases fell outside of the 95% binomial CI after $n = 401$. In the conditions with the smallest sample size ($n = 101$), there were slight differences in the size of the biases between model forms.

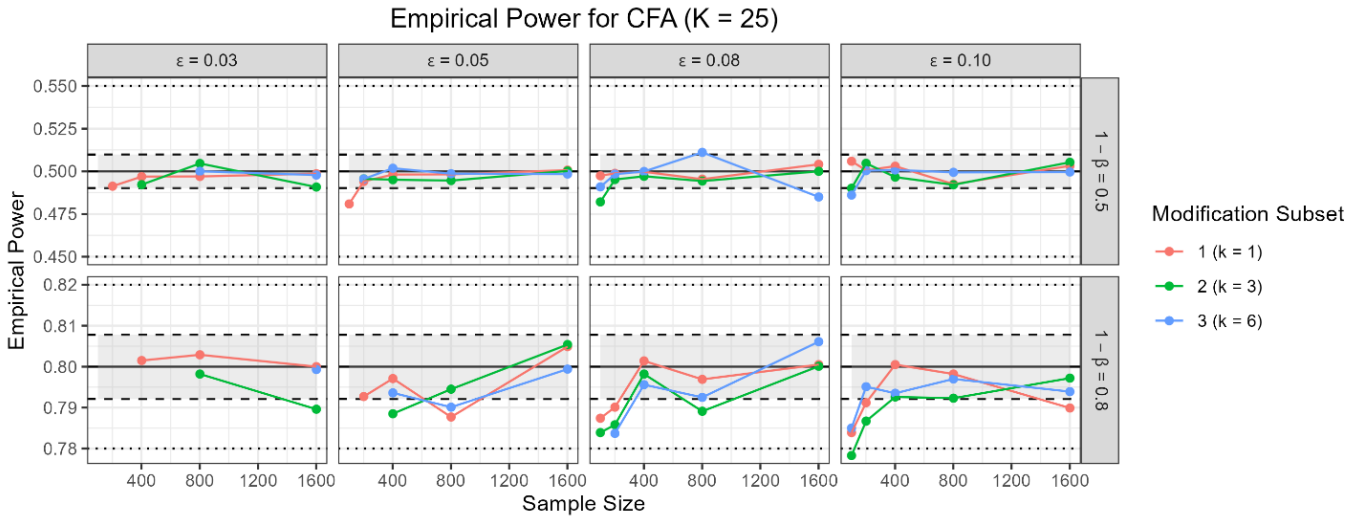
In the CFA model form condition ($K = 25$), the only practically significant bias was observed for the test of modification subset 2 ($k = 3$), when $n = 101$ and $\epsilon = 0.10$ at 0.8 nominal power. It had an absolute relative size of around 12%. As mentioned in the asymptotic power analysis section, many of the tests in the CFA model form condition were underpowered at low sample size and small approximation error. Extrapolating the general trends in the analysis noted above, we would expect the tests in these conditions to have the largest biases. However, because the tests were underpowered, they were omitted from the analysis, and hence there are no results to report.

Similarly, the LVPM model form condition ($K = 86$) only had practically significant biases for 0.8 nominal power. However, there were more practically significant biases in the CFA condition, and biases were generally larger in size, with the largest bias being around 18%. At $n = 101$ and $\epsilon = 0.05$, the tests of modification subsets 1 ($k = 1$) and 2 ($k = 6$) had practically significant biases, with the test of larger subset having a larger bias. At $n = 101$ and $\epsilon = 0.08$, practically significant biases of similar size occurred on tests of modification subsets 1 ($k = 1$) and 3 ($k = 15$). Finally, only the test of modification subset 1 was practically significant at $n = 101$ and $\epsilon = 0.10$. As with the CFA model form condition, the tests were underpowered in conditions where we would expect the largest biases. Because they were omitted, we do not have results in these conditions.

The practically significant biases in the REPCFA model form condition ($K = 133$) were both more numerous and larger than those in the other two model form conditions. The largest absolute relative bias was 38%. The test of modification subset 1 ($k = 1$) had practically significant biases at $n = 101$ for $\epsilon = 0.03, 0.05$, and 0.08 at 0.8 nominal power. The test of modification subset 2 ($k = 9$) had practically significant biases in the $n = 101$ and $\epsilon = 0.03$ condition at 0.5 nominal power. At 0.8 nominal power, the test of this subset had practically significant biases at $n = 201$ and $\epsilon = 0.03$, and $n = 101$ when $\epsilon = 0.05$ and 0.08. Finally, the test of the third modification subset ($k = 21$) was practically significant at $n = 101$ in the $\epsilon = 0.05$ and 0.08 conditions at 0.5 nominal power, and in all approximation error conditions at 0.8 nominal power. Across all approximation error conditions, the biases at $n = 101$ increased in size as the size of the modification subset increased.

Figure 10

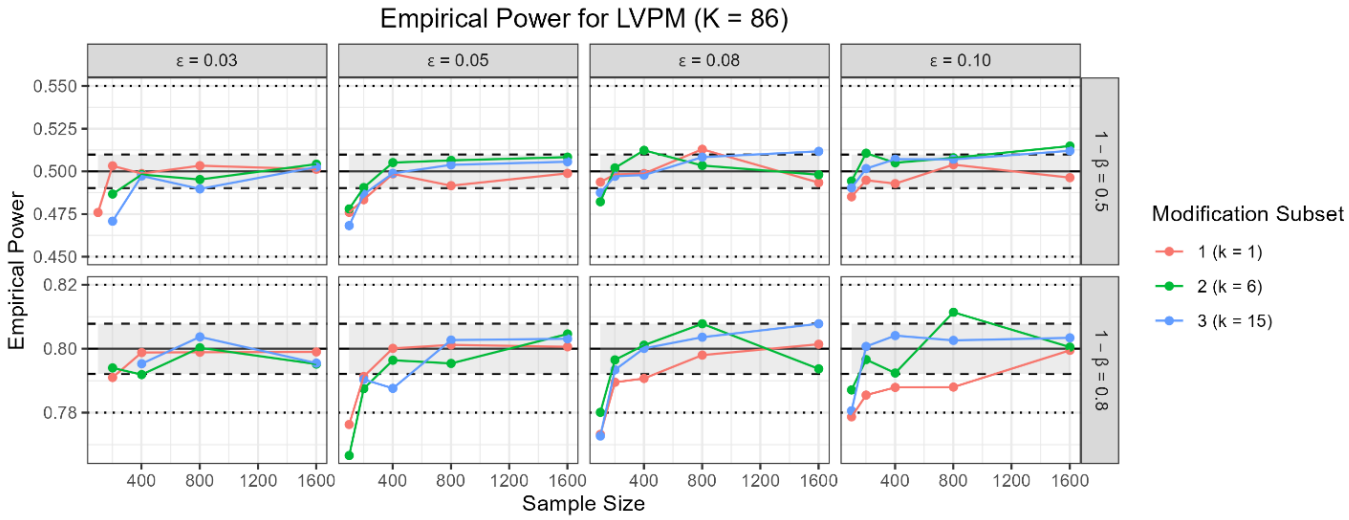
Empirical power for Model 1 (CFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Figure 11

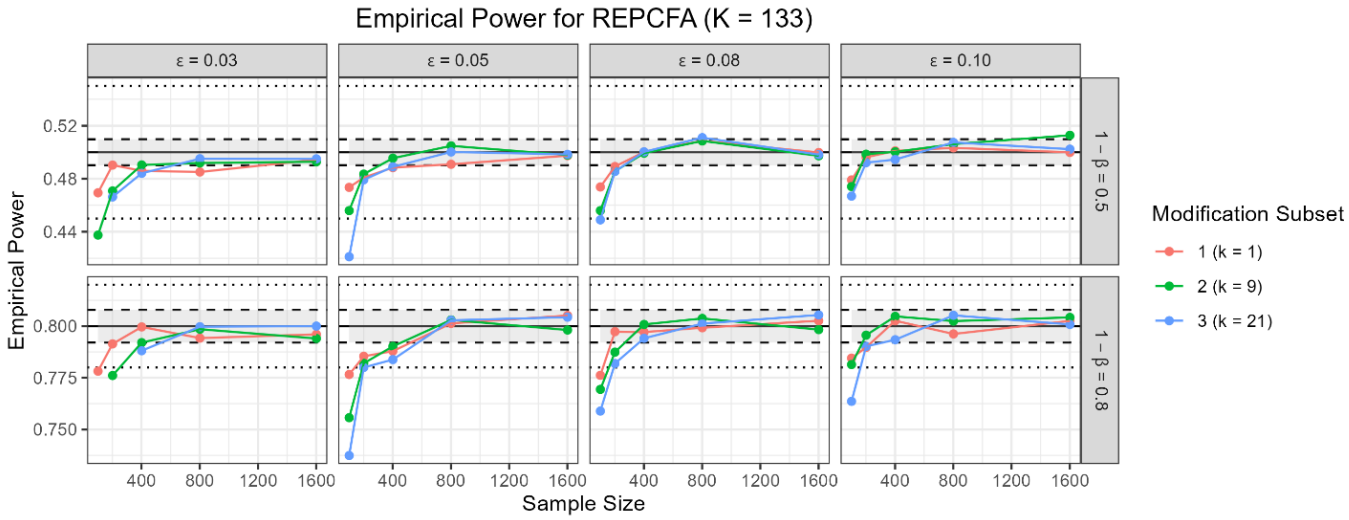
Empirical power for Model 2 (LVPM)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Figure 12

Empirical power for Model 3 (REPCFA)



Note. K = model degrees of freedom, k = modification subset degrees of freedom, ϵ = RMSEA (approximation error factor), $1 - \beta$ = theoretical power.

Table 2

Effects of manipulated factors on finite sample bias in empirical power

Source	df	F	p	partial η^2
Model Form (Form)	2	7.55	< 0.001	0.12†
Modification Subset (Subset)	2	0.37	0.692	0.01
Sample Size (n)	4	112.88	< 0.001	0.81††
Form × Subset	4	5.12	< 0.001	0.16†
Form × n	8	5.55	< 0.001	0.29††
Subset × n	8	3.39	0.002	0.20†
Form × Subset × n	16	1.59	0.082	0.19†
Approximation Error (ϵ)	1	15.81	< 0.001	0.13†
$\epsilon \times n$	4	3.99	0.005	0.13†

Note. All factors are modeled as categorical variables, with separate indicators for each non-reference level, except Approximation Error (ϵ), which was modeled as a linear relation. Daggers denote the Cohen et al. (2003) partial η^2 thresholds for multiple regression delineating small-medium and medium-large sized effects, respectively,

$^{\dagger}0.09 \leq \eta^2 < 0.26$, $^{\dagger\dagger}0.26 \leq \eta^2$

Chapter 5: Discussion

Evenly Dispersed Approximation Error

While perfect model fit is unattainable, useful models provide a simplified but sufficiently accurate approximation of reality. In pursuit of such models, some misspecification is inevitable, but minor deviations are often accepted for the sake of interpretability and theoretical justifiability. Global fit evaluation assesses whether a model's overall misspecification is acceptably small, but it risks overlooking localized sources of misfit. Local fit evaluation, though valuable for detecting misspecifications, fails to contextualize the misfit within the fit of the other model components. Crucially, neither approach sufficiently defines a heuristic for acceptable local approximation error given acceptable global approximation error.

To address this gap, the current work proposed *evenly dispersed approximation error* as an additional criterion for model acceptability. Since misspecification is unavoidable, the most useful model balances fit, parsimony, interpretability, and theoretical justification (MacCallum, 2001). No single modification should disproportionately improve fit; otherwise, that modification would yield a superior model. It follows, then, that this condition requires local approximation error to be roughly uniform.

Although the concept provides a useful framework for evaluating local fit, two key concerns remain. First, the approach relies on the assumption that the initially hypothesized model provides a close enough approximation of the population that local misfit meaningfully reflects the validity of the corresponding hypotheses. However, conventional global fit thresholds (e.g., RMSEA < 0.05, SRMR < 0.08; Hu & Bentler, 1999) may not ensure this assumption holds. Second, although evenly dispersed misfit is intuitive, alternative conceptualizations may exist that better capture the spirit of acceptable local approximation error.

For example, it seems reasonable to set absolute limits on local misspecification independent of global misfit. However, this approach would face several challenges. A fundamental difficulty lies in developing a local analog to the close approximation of data principle that informs global fit thresholds, primarily in defining what aspect of the data would be considered local. Furthermore, because multiple hypotheses can be responsible for the same misfit (MacCallum, 1986; MacCallum et al., 1992), it would be necessary to formally specify how to decompose the misfit when evaluating local approximation error for each hypothesis separately. Finally, the interpretability of local fit depends on the model estimated parameters and moments being close to their true population values. Thus, local fit evaluation may not be meaningful without ensuring that global fit is adequate.

Operationalizing Misfit Dispersion

To operationalize the concept of evenly dispersed approximation error, I extended a powerful property from sampling error. When the model is correctly specified, such that it would have exact fit to the population, its misfit with a sample from that population is necessarily nonsystematic and unavoidable. For any modification, the proportion of misfit it would resolve is asymptotically beta distributed given the global fit. Notably, the proportion is also asymptotically independent of the global fit, depending only on the degrees of freedom for the model and modification. Taking this property as the defining characteristic evenly dispersed misfit implies that the local misfit in models meeting this criterion for good fit is controlled within well-defined limits, regardless of the presence or magnitude of global approximation error.

The fundamental limitation of this definition lies in the assumed degree of conceptual symmetry between sampling and approximation error. Such parity is not unprecedented in SEM fit evaluation. Take, for example, the 0.05 threshold for RMSEA (Steiger, 2016; Steiger & Lind,

1980; Browne & Cudeck, 1993). For a sample with $n = 401$ observations, this value implies that the global misfit is comprised of roughly even proportions of both types of error. Specifically, the expected value of the estimate of the noncentrality parameter $E(\hat{\delta}) = T - K$ is equal to the model degrees of freedom K, which is also the expected value of the global LRT fit statistic T under the hypothesis of exact fit. While the threshold was not intentionally developed for this reason, it implicitly extends the interpretive scale for sampling error to approximation error. This extension parallels the approach to operationalization taken in this work.

In discussing the interpretational framework for RMSEA developed by Wu and Browne (2015), Satorra (2015) and Maydeu-Olivares (2017) critiqued the treatment of approximation error as random. The primary criticism in both articles was the lack of evidence for the general existence of a superpopulation from which a theoretically random approximation error would be sampled. Unfortunately, I cannot satisfactorily resolve this concern directly. However, as I argued in the theoretical derivation, in the absence of an inherent scale for the unevenness of deterministic approximation error, measuring it against the unevenness sampling error is, at the very least, intuitive if not consistent with our expectations.

One minor criticism in Satorra (2015) concerned the limited distributional form assumed for approximation error in the framework. In contrast, the result in the current work holds for any elliptical distribution, which is a much larger class of distributions than that used in Wu & Browne (2015). The defining feature of this class is elliptical symmetry, which, in simple terms, implies a uniform distribution on all the ways a model can be misspecified. Yet, it is infinitely more restrictive than if no distribution had been assumed approximation error at all.

This dilemma bears resemblance to the famously divisive discussion on noninformative priors in Bayesian inference. As Gelman. (2006) argued, the notion is trivially paradoxical.

Noninformative priors are, in fact, highly informative, particularly when analyzing small samples. For example, a flat prior on a positive parameter implicitly favors extreme values, placing infinite mass on every right-unbounded interval. Thus, from this perspective, noninformative priors require the same level of justification as informative priors.

In the present context, alternative operationalizations for evenly dispersed error would entail some degree of concentrated misfit. Recall that, given any orthogonal basis of modifications, the misfit proportions are asymptotically distributed as $Dirichlet\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$ and independent of global fit (Fang & Zhang, 1990). The symmetry in the concentration parameters results in identical marginal distributions, which is one of the necessary conditions for evenly dispersed misfit. The second condition, independence of orthogonal basis, would not hold if the parameters had any other value than $1/2$.

As a tangible example, it may be reasonable to operationalize evenly dispersed misfit such that less misfit is concentrated on particularly focal hypotheses than hypotheses required for model identification. Additionally, the concentration could depend on global fit, such that worse fitting models would require less concentration on focal hypotheses than better fitting models. For instance, the asymptotic null distribution of the misfit proportion statistic could be $Beta\left(\frac{k}{2}, \frac{K-k}{2}; \delta\left(\frac{k}{K}\right), \delta\left(1-\frac{k}{K}\right)\right)$. In practical terms, this formulation would arise from the assumption that all local approximation error is exactly proportional to the proportion of the model degrees of freedom contained in the modification subset k/K. As with most assumptions in statistics, these decisions are unfalsifiable empirically, and must therefore be justified with context-relevant theory or an appeal to mathematical convenience in service of a practical tool. For the current work, I can only offer the latter.

On the topic of the elliptical symmetry, it should be reiterated that this is a property of the asymptotic distribution of misfit under exact fit. The evenness of sampling error may manifest quite differently in finite samples, especially as the discrepancy between model-implied and sample moments increases. A more sophisticated understanding of differential geometry may inform a rigorous operationalization of evenly dispersed misfit that captures non-local features of the discrepancy function.

A final consideration concerns the operationalization's reliance on regularity conditions that ensure the asymptotic normality of moment residuals under exact fit. If violations to these assumptions were suspected, such as nonnormality, alternative test statistics may be necessary. The asymptotically distribution-free (ADF) estimator (Browne, 1984) or robust scaled statistics (Satorra & Bentler, 1988, 1994; Yuan & Bentler, 2000) offer potential solutions, though each carries its own theoretical and computational trade-offs. Lack of robustness would likely compound with finite sample errors (Moshagen et al., 2012; Yuan & Bentler, 2006; Yuan et al, 2015). While the simulation study in the following section empirically evaluates convergence under normal distribution assumptions, future research should examine performance with nonnormal data and nonlinear parameterizations.

Simulation Studies

The simulation studies aimed to determine sufficient sample size for convergence of the misfit proportion statistic and investigate sources of finite sample bias for the misfit proportion statistics. The results revealed relatively small bias overall. Even at the smallest sample size ($n = 101$), before most statistics had converged, the maximum observed relative bias was 38%. At $n = 401$ and above, the largest relative bias was merely 17%, with most estimates falling well below the practically significant 10% threshold. These findings suggest that samples with at least

$n = 201$ observations are sufficient for accurate approximation of both the asymptotic null and alternative distributions. This is consistent with previous empirical evaluations of convergence of the likelihood ratio test (LRT) statistic (Yuan, 2005).

Certain underpowered conditions were necessarily excluded from empirical power analysis, including many that fell below the $n = 201$ threshold. While we might reasonably assume these would similarly fail to converge to their asymptotic levels, such extrapolation becomes less justified for the second and third CFA modification subsets, which remained underpowered even at $n = 401$ and 801, respectively. However, this limitation carries minimal practical significance as these use cases would be discouraged due to being underpowered regardless of convergence.

Regression analyses in both studies demonstrated the impact of the model size effect on the finite sample distributions of the misfit proportion statistic. As model degrees of freedom (K) increased, bias in the empirical distributions grew increasingly negative. If driven primarily by the denominator LRT statistic, this result would align with findings on model size effects in LRT statistics, which identified an increasingly positive bias with increasing K (Moshagen et al., 2012; Shi et al., 2019). This conclusion is plausible given the denominator having more degrees of freedom than the numerator and should thus be more impacted by the model size effect. However, this is complicated by our findings on the biasing effects of increasing modification subset degrees of freedom (k). Although much more modest than the trend with K, bias also decreased with increasing k in both studies, when the opposite would be expected. This may suggest that the effects governing finite sample bias behave unexpectedly when taking the difference of LRT statistics from nested models. Furthermore, there may be specific characteristics of the constraints in modification subsets that modify the general model size

effect. Future research into the asymptotic properties of LRT difference statistics is required to clarify the underlying mechanisms relating bias to k.

Both studies found evidence of a substantial approximation error effect. Across model form conditions, increasing approximation error corresponded to increasingly positive bias, partially offsetting the negative bias from the model size effect at smaller sample sizes. This pattern mirrors the biasing effect of approximation error for the alternative distribution of LRT (Yuan et al., 2007). In Yuan (2005) the upper tail of the distribution appeared to have an increasingly large downward bias as approximation error increased. Similar to the model size effect, this phenomenon could propagate through the denominator statistic, potentially explaining the upward bias observed in our simulations.

The opposing directionality of bias from model size effects and approximation error effects may indicate independent underlying mechanisms that converge differently with increasing sample size. For the two smallest models (CFA and LVPM), greater approximation error led to elevated false positive rates in the first study. In contrast, the two sources of bias appeared to cancel out at large approximation error for the largest model (REPCFA). This raises the possibility that the statistics in the REPCFA condition may not be fully converged to their asymptotic distribution by $n = 201$. If this is the case, then future research is required to disentangle these two effects to ensure the sample size requirements are consistent.

In summary, the simulation studies indicated that both the null and alternative empirical distributions are sufficiently close to their asymptotic equivalents around a sample size of $n = 201$. This finding carries significant practical implications: researchers working with sample sizes can (1) use the asymptotic central beta distribution to test the hypothesis of evenly

dispersed misfit using the misfit proportion statistic and (2) use the asymptotic doubly noncentral beta distribution for *a priori* power analysis when design studies.

However, due to the limited scope of the simulation studies, these conclusions may not generalize. Given the substantial bias reported by Moshagen et al. (2012) for models larger than those examined here, testing more complex models would likely require samples exceeding $n = 201$. Additionally, extrapolating the approximation error effect beyond the largest value in this study ($\epsilon = 0.10$), the bias may become similarly untenable at $n = 201$. However, this degree of misfit would exceed most generally accepted thresholds for approximation error (Browne & Cudeck, 1993; Hu & Bentler, 1999). Thus, if the misfit proportion test is used exclusively for good fitting models, this should not be a problem in practice.

While the distributions, model forms, and modification subsets used to generate the simulated data resemble conditions commonly encountered in practice, they do not fully capture the diversity of contexts in which SEM is applied. The data and models were generally consistent with the assumptions of the normal theory maximum likelihood (NT-ML) estimator (Yuan & Bentler, 2006). When these assumptions are violated, LRT may be biased or have an entirely different asymptotic distribution entirely. rendering the asymptotic distribution of the misfit proportion statistic indeterminate.

The simulations were limited to models with linear relations and constraints. However, frameworks have been developed for SEM that model nonlinear relations (Harring et al., 2012; Lee & Zhu, 2002), latent variable interactions (Klein & Moosbrugger, 2000), nonlinear constraints (Savalei & Kolenikov, 2008), and inequality constraints (van de Schoot et al., 2010). Furthermore, frameworks have been developed for specially structured data, such as multiple groups (Jöreskog, 1971; Sörbom, 1974), multilevel structures (Asparouhov & Muthén, 2008),

and mixture models (Muthén, 2002). While the NT-ML LRT statistics for these frameworks remain asymptotically χ^2-distributed for normal data, their performance in finite samples remains largely unstudied, making predictions about the misfit proportion statistic's behavior uncertain.

Additionally, the simulations were limited to continuous, normally distributed data. Conveniently, the results from these studies may be extended to nonnormal data by using alternative LRT statistics in the place of the NT-ML LRT statistics. Browne's (1984) asymptotically distribution-free (ADF) estimator yields asymptotically χ^2-distributed LRT statistic, regardless of data distribution. More commonly, the NT-ML LRT statistics are rescaled to achieve asymptotic χ^2 distributions for covariance structure models (Satorra & Bentler, 1988), mean structure models (Satorra, 1992; Satorra & Bentler, 1994; Yuan & Bentler, 1999), and multigroup models (Satorra, 2000). However, the asymptotic robustness of these scaled statistics has primarily been established for models with exact fit and elliptically distributed data (Browne, 1984; Shapiro & Browne, 1987). Under more general nonnormality, scaled statistics may fail to follow their asymptotic χ^2 distributions, with accuracy potentially worsening as sample size increases (Bentler & Yuan, 1999; Yuan & Bentler, 1998). No empirical research has examined scaled statistics under model misspecification, though results for unscaled NT-ML LRT (Yuan et al., 2005) suggest their asymptotic distributions may deviate from the noncentral χ^2.

When their assumptions hold, alternative estimators and scaling approaches yield LRT statistics that are asymptotically χ^2-distributed, implying the asymptotic beta distribution for the misfit proportion statistic should remain valid. However, these alternatives generally require larger samples than their NT-ML counterparts, with ADF-based statistics requiring prohibitively large samples (Hu et al., 1992). The misfit proportion statistic would likely share these regularity

conditions and sample size requirements. The impact of assumption violations on these alternative LRT statistics, and hence on the misfit proportion statistic, remains understudied. Further research into finite sample behavior and its consequences for the misfit proportion statistic is necessary.

The Case for Misfit Dispersion Evaluation

When synthesizing global and local fit evaluation, the current SEM practice suffers from an inherent inconsistency. The traditional approach recommends approximate global fit measures while, until relatively recently, only exact fit tests, such as Likelihood Ratio difference tests or Lagrange Multiplier tests, existed for local fit evaluation. This creates a mixed standard. Researchers tolerate approximation error in total model misfit yet expecting none in the constituent hypotheses, despite the former being an aggregate of the latter. Such inconsistency weakens the conclusions drawn from overall fit evaluation.

As an approach for *approximate* local fit evaluation, the proposed misfit dispersion framework, including the misfit proportion test analyzed here, directly addresses this inconsistency. Approximate local fit methods extend the well-established logic of approximate global fit measures to local fit evaluation. Consequently, this approach to local fit evaluation can be unified with global fit evaluation under a single, consistent logic of approximate fit.

From a practical perspective, methods for approximate local fit evaluation generally establish a higher threshold for unacceptable misfit that scales with sample size. In effect, this threshold more consistently privileges the *a priori*, theory-based specification of the original hypothesized model compared to the heightened sensitivity of exact fit tests. Specifically, within the context of specification search (MacCallum, 1986; MacCallum et al., 1992), the more stringent criteria of approximate local fit measures lead to earlier termination of the search

process. This reduces the risk of overfitting to the sample, ultimately yielding more parsimonious models.

The misfit dispersion framework joins recent developments in approximate local fit evaluation, such as a local implementation of RMSEA ($RMSEA_D$; Beribisky & Hancock, 2023) and Opdyke intervals for correlation residuals (McNeish, 2025). It shares many strengths with these alternatives but also offers distinct advantages. Both the misfit dispersion framework and Opdyke intervals possess inherently meaningful scales for measuring misfit. These scales derive from a probabilistic operationalization of the expected variability of misfit under some definition of acceptable model misspecification. Furthermore, both approaches utilize representations of local misfit that respect its multidimensional nature and its aggregation into global misfit.

In contrast, $RMSEA_D$ (Beribisky & Hancock, 2023) lacks such an inherently meaningful scale. As a direct adaptation of the global RMSEA index, $RMSEA_D$ values are given meaning by intuitive or simulation-based thresholds, which depend on underexamined assumptions (Steiger, 2007). The $RMSEA_D$ measure also does not treat local fit accurately. It effectively treats the tested submodel as a self-contained model, neglecting overlapping misfit from the rest of the mode. Literature notes its poor performance with low degrees of freedom (Kenny et al., 2015; Shi et al., 2022). Under these conditions, $RMSEA_D$ can be overpowered compared to the misfit proportion test, sometimes exhibiting even greater sensitivity than exact fit tests.

Compared to Opdyke intervals (McNeish, 2025), the misfit dispersion framework also has notable benefits. Opdyke intervals are limited to inferences on correlation residuals, which may not correspond directly to substantive hypotheses. As a result, drawing conclusions about misspecifications that are meaningful to the model specification can be difficult. This limitation also makes Opdyke intervals less suitable for exhaustive exploratory analysis. Conversely, the

misfit dispersion framework allows the evaluation of arbitrary subsets of constraints. This flexibility facilitates both confirmatory hypothesis testing and exhaustive exploration of misfit.

The theoretical foundation of the misfit dispersion framework constitutes another strength. The framework enables generalization to any discrepancy measure. This flexibility is particularly valuable when using the measure underlying the chosen global fit evaluation. For the simulations conducted in this work, the Maximum Likelihood discrepancy function, defining the parameter estimator, was used. This gave it particular synergy with RMSEA. However, future efforts could extend this approach to align with other global fit indices, such as SRMR (Maydeu-Olivares, 2017). Furthermore, the framework was developed from a novel characterization of a property of global misfit, the hypothesis of evenly dispersed misfit. Because of this, the misfit dispersion framework could form a basis for refining global fit evaluation. The next subsection briefly outlines some approaches for testing and exploring the hypothesis of globally even dispersion of misfit.

Future Directions

The ambiguity of subset selection and non-rejection for approaches to local fit evaluation like the misfit proportion test highlights the need for principled method for omnibus evaluation. Given the prominence of global fit indices in SEM practice, researchers would benefit from global tests of evenly dispersed misfit that can be computed directly from model output without additional specification. Though developing such methods extends beyond the scope of the current work, the following outlines future directions to address this gap based on the theory of misfit dispersion.

Tests of elliptical and spherical symmetry provide a promising basis for developing an omnibus test of misfit dispersion (Kariya & Eaton, 1977; King, 1980). When treating the

Lagrange multipliers (LMs) $\hat{\boldsymbol{\lambda}}$ of the model as a sample from a normal distribution with covariance matrix $\boldsymbol{\Sigma}$, the statistic $U = \hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}/\hat{\boldsymbol{\lambda}}'\boldsymbol{\Omega}^{-}\hat{\boldsymbol{\lambda}}$ tests the hypothesis that $\boldsymbol{\Sigma} = \mathbf{I}$ against the alternative that $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$. When $\boldsymbol{\Omega}$ is defined to be the asymptotic covariance matrix of the LMs (see Yuan and Bentler, 2006), then U measures how consistent $\hat{\boldsymbol{\lambda}}$ is with elliptical symmetry compared to spherical symmetry.

Note that the square of the standardized LMs defines the modification index (MI) test statistic $T_{MI} = (n-1)\ell_i^2$, where the standardized LM is given by $\ell_i = \hat{\lambda}_i/\sqrt{\Omega_{ii}}$ (Yuan and Bentler, 2006). Using $\boldsymbol{\ell}$ in place of $\hat{\boldsymbol{\lambda}}$ and the asymptotic LM correlation matrix $\boldsymbol{\Omega}_\ell$ in place of $\boldsymbol{\Omega}$, the statistic $U_\ell = \frac{(n-1)\boldsymbol{\ell}'\boldsymbol{\ell}}{(n-1)\boldsymbol{\ell}'\boldsymbol{\Omega}_{\boldsymbol{\ell}}^{-}\boldsymbol{\ell}}$ has a numerator equal to the sum of MIs and a denominator equal to the LM estimate of the global model fit statistic. Thus, dividing U_ℓ by the number of MIs m, we can see that the statistic U_ℓ/m is equivalent to the average MI proportion.

Conceptually, if misfit were evenly dispersed, it would not be particularly aligned with any subset of model hypotheses. As a result, the average MI proportion would be close to $1/K$, which is the expected proportion of misfit for a modification subset with $k = 1$. If misfit were more aligned with a subset of model hypotheses than others, the average MI proportion would be larger than $1/K$.

Preliminary exploration of U, U_ℓ, and U_ℓ/m has been encouraging. Using Monte Carlo simulation to generate samples under the null and various alternative hypotheses, I found that unevenly dispersed error could be reliably detected by testing for a large average MI proportion. The power is slightly lower than the confirmatory misfit proportion test, as would be expected, and notably decreases as the modification subset in which the misfit is concentrated under the alternative hypothesis grew larger.

Exploratory applications of the misfit proportion statistic may provide a fruitful alternative to omnibus statistics. The theory that underlies Lagrange Multipliers is based on an approximation of discrepancy-based SEM estimators that represents refitting a modified model as a linear regression of the moment residuals onto the model constraints (Shapiro, 1986). As a result, we can apply variable selection methods to identify a constraint subset for the misfit proportion test.

One potential approach involves stepwise constraint selection to identify a subset with highly concentrated misfit. Researchers could iteratively add or remove constraints until no remaining constraint significantly alters the disproportionality of the misfit concentrated in the subset. Alternatively, principal components analysis of the asymptotic LM covariance or correlation matrix could provide a more qualitative approach. By projecting moment residuals onto these components and examining squared residual loadings, researchers could assess their consistency with a $Dirichlet\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$ distribution. Disproportionately large squared loadings on components explaining substantial LM variance would suggest systematic association between misfit and constraints, and thus uneven misfit dispersion. Notably, this is inherently connected to the elliptical symmetry statistic detailed above, which essentially measures the correlation between squared residual loadings and the variance explained by the corresponding component. Future research should systematically evaluate these proposed omnibus and exploratory methods.

Conclusion

This dissertation has established a fundamental refinement for evaluating structural equation models: the hypothesis of evenly dispersed approximation error. Moving beyond the

traditional reliance on aggregate global fit, I argued that a well-specified model must demonstrate that its constituent hypotheses *each* provide a close approximation to reality, not just in aggregate. To operationalize this principle, I introduced the misfit dispersion framework and formally derived the asymptotic properties of the misfit proportion statistic, defining an inferential test for evenly dispersed misfit. Simulation studies demonstrated the performance of the test in finite samples, confirming that these asymptotic distributions provide reliable inferences for moderately sized samples (around 200 observations).

Integrating the evaluation of misfit dispersion into standard SEM is straightforward. Implementation can be as simple as replacing traditional likelihood ratio difference tests with the misfit proportion test. Conveniently, this does not require any additional computation; the statistics used in the test are already provided by standard SEM software. With such a low barrier to entry, this approach is an accessible addition to the approximate local fit evaluation toolbox.

More profoundly, it resolves the inherent incompatibility between approximate global fit and exact local fit evaluation. It facilitates a unified criterion for model acceptability, requiring the global approximation error be acceptably small *and* evenly dispersed across constituent hypotheses. This dual requirement ensures the validity of inferences drawn from all parts of the model.

Future work will focus on developing omnibus and exploratory tools for global misfit dispersion evaluation, further enhancing the framework's accessibility and utility. With its strong theoretical basis, empirical support, and practical accessibility, the misfit dispersion framework represents a vital new tool for enhancing fit evaluation in structural equation modeling.

Appendix A. Misspecification Generation Method

Populations were generated using a method adapted from Lai (2019). In the original presentation, the author described a process for obtaining moment residuals $\boldsymbol{\rho} = [\mathbf{t}', \text{vech}(\mathbf{E})']'$ such that

$$\widetilde{\boldsymbol{\theta}} = \underset{\mathbf{h}(\boldsymbol{\theta})=\mathbf{0}}{\text{argmin}}\, F\left(\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}}) + \boldsymbol{\rho}, \boldsymbol{\beta}(\boldsymbol{\theta})\right)$$

$$F\left(\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}}) + \boldsymbol{\rho}, \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})\right) = F_0$$

for a given $\widetilde{\boldsymbol{\theta}}$ and F_0. Because $\widetilde{\boldsymbol{\theta}}$ minimizes the discrepancy function, $\boldsymbol{\rho}$ satisfies the first order condition

$$\boldsymbol{\Delta}_f \mathbf{V} \boldsymbol{\rho} = \mathbf{0}.$$

The infinitely many solutions of the condition have the form

$$\boldsymbol{\rho}(\mathbf{x}) = \left(\mathbf{I} - \left(\boldsymbol{\Delta}_f \mathbf{V}\right)^{-} \boldsymbol{\Delta}_f \mathbf{V}\right)\mathbf{x} = \mathbf{P}\mathbf{x}$$

for arbitrary $\mathbf{x}$, where $\left(\boldsymbol{\Delta}_f \mathbf{V} \boldsymbol{\rho}\right)^{-}$ denotes the generalized inverse. Given some initial value $\mathbf{x}^{(0)}$, a solution $\mathbf{x}$ that achieves the desired discrepancy F_0 can be obtained by solving $F\left(\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}}) + \boldsymbol{\rho}(\mathbf{x}), \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})\right) - F_0 = 0$ numerically.

To generate a $\boldsymbol{\rho}(\mathbf{x})$ that corresponds to evenly dispersed approximation error, I developed the method to sample $\mathbf{x}$ in a more principled manner. Notice that, based on the second-order approximation of F (Shapiro, 1986), any solution $\mathbf{x}$ approximately satisfies

$$\mathbf{x}'\mathbf{P}\mathbf{V}\mathbf{P}\mathbf{x} \approx F_0.$$

This condition implies that solutions have the approximate form $\mathbf{x} \approx \mathbf{L}\mathbf{z}$ for some $\|\mathbf{z}\|_2 = \sqrt{F_0}$, where $\mathbf{L}'\mathbf{P}\mathbf{V}\mathbf{P}\mathbf{L} = \mathbf{I}$. Because $\text{rank}(\mathbf{P}) = K$, the model degrees of freedom, $\mathbf{z}$ contains K entries. Given independently distributed $\xi_i \sim N(0,1)$ for $i = 1, \ldots, K$, we can define an initial value

$$\mathbf{x}^{(0)} = \frac{\sqrt{F_0}}{\|\boldsymbol{\xi}\|_2}\mathbf{P}\boldsymbol{\xi}$$

which is distributed uniformly on the surface of the Kdimensional hypersphere with radius $\sqrt{F_0}$. Because the spherical uniform distribution is elliptically symmetric, the approximation error resulting from $\boldsymbol{\rho}(\mathbf{x})$ is definitionally evenly dispersed.

Misfit was targeted into a set of constraints by defining $\mathbf{Q} = (\boldsymbol{\Delta}_m\mathbf{V})^{-}\boldsymbol{\Delta}_m\mathbf{V} - \left(\boldsymbol{\Delta}_f\mathbf{V}\right)^{-}\boldsymbol{\Delta}_f\mathbf{V}$ where $\boldsymbol{\Delta}_m$ is the Jacobian matrix for both the free parameters and the constraint subset. Whereas $\mathbf{P}$ projects approximation error over all constraints, $\mathbf{Q}$ targets the constraint subset specifically. Thus, we can obtain $\boldsymbol{\rho}(\mathbf{u})$ for which $\mathbf{u}'\mathbf{Q}\mathbf{V}\mathbf{Q}\mathbf{u} \approx F_0\pi_{\text{diff}}$ and $\boldsymbol{\rho}(\mathbf{w})$ for which $\mathbf{w}'(\mathbf{P}-\mathbf{Q})\mathbf{V}(\mathbf{P}-\mathbf{Q})\mathbf{w} \approx F_0(1-\pi_{\text{diff}})$, for a given π_{diff}. As a result, $\boldsymbol{\beta}(\tilde{\boldsymbol{\theta}}) + \left(\boldsymbol{\rho}(\mathbf{u}) + \boldsymbol{\rho}(\mathbf{w})\right)$ will have the target discrepancy F_0 with $\boldsymbol{\beta}(\tilde{\boldsymbol{\theta}})$ and the desired proportion π_{diff} will be attributable to the selected constraint subset.

Appendix B. Simulation Study 1 Results

Table B1. *Empirical false positive rates for Model 1: (CFA, $K = 25$)*

		Modification Subset 1 ($k = 1$)			Modification Subset 2 ($k = 3$)			Modification Subset 3 ($k = 6$)		
RMSEA	**Sample Size**	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.00	101	0.104	0.052	0.010	0.103	0.052	0.010	0.104	0.052	0.011
	201	0.102	0.051	0.010	0.101	0.050	0.010	0.100	0.050	0.010
	401	0.102	0.051	0.011	0.101	0.051	0.010	0.101	0.051	0.010
	801	0.100	0.050	0.009	0.101	0.050	0.010	0.098	0.049	0.010
	1601	0.101	0.050	0.010	0.100	0.049	0.010	0.100	0.049	0.010
0.03	101	0.100	0.050	0.011	0.104	0.053	0.011	0.105	0.054	0.011
	201	0.102	0.051	0.010	0.101	0.051	0.010	0.102	0.050	0.010
	401	0.102	0.049	0.010	0.102	0.050	0.010	0.102	0.052	0.011
	801	0.099	0.049	0.010	0.103	0.052	0.010	0.099	0.050	0.010
	1601	0.101	0.049	0.010	0.101	0.050	0.011	0.100	0.052	0.010
0.05	101	0.105	0.054	0.011	0.104	0.055	0.011	0.107	0.055	0.011
	201	0.105	0.054	0.011	0.102	0.050	0.011	0.106	0.052	0.011
	401	0.100	0.051	0.010	0.103	0.052	0.010	0.102	0.051	0.010
	801	0.099	0.049	0.009	0.102	0.051	0.010	0.103	0.051	0.011
	1601	0.102	0.051	0.010	0.101	0.050	0.009	0.100	0.049	0.010
0.08	101	0.108	0.055	0.012	0.107	0.055	0.011	0.111	0.056	0.011
	201	0.106	0.054	0.011	0.104	0.052	0.011	0.103	0.051	0.010
	401	0.100	0.051	0.011	0.102	0.051	0.011	0.102	0.052	0.011
	801	0.102	0.052	0.011	0.101	0.050	0.010	0.100	0.050	0.010
	1601	0.102	0.051	0.010	0.099	0.050	0.010	0.101	0.050	0.010
0.10	101	0.108	0.057	0.011	0.108	0.055	0.012	0.111	0.055	0.010
	201	0.104	0.053	0.011	0.105	0.053	0.011	0.105	0.052	0.011
	401	0.107	0.054	0.010	0.102	0.050	0.009	0.102	0.050	0.010
	801	0.101	0.050	0.010	0.101	0.051	0.010	0.102	0.052	0.010
	1601	0.102	0.052	0.010	0.104	0.052	0.010	0.101	0.052	0.011

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and α = theoretical false positive rate.

Table B2. *Empirical false positive rates for Model 2: (LVPM, $K = 86$)*

		Modification Subset 1 ($k = 1$)			**Modification Subset 2 ($k = 6$)**			**Modification Subset 3 ($k = 15$)**		
RMSEA	**Sample Size**	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.00	101	0.097	0.048	0.009	0.099	0.049	0.010	0.097	0.050	0.011
	201	0.097	0.048	0.010	0.099	0.049	0.010	0.100	0.049	0.010
	401	0.099	0.049	0.010	0.099	0.049	0.009	0.099	0.050	0.009
	801	0.100	0.051	0.010	0.099	0.050	0.010	0.100	0.050	0.010
	1601	0.098	0.049	0.011	0.102	0.052	0.011	0.099	0.050	0.010
0.03	101	0.097	0.048	0.010	0.100	0.050	0.010	0.099	0.050	0.011
	201	0.100	0.049	0.009	0.102	0.051	0.010	0.100	0.050	0.010
	401	0.101	0.049	0.009	0.101	0.051	0.010	0.101	0.051	0.010
	801	0.100	0.048	0.010	0.102	0.051	0.011	0.101	0.049	0.009
	1601	0.100	0.051	0.011	0.100	0.050	0.010	0.101	0.051	0.010
0.05	101	0.101	0.050	0.009	0.102	0.052	0.010	0.100	0.051	0.010
	201	0.101	0.052	0.010	0.104	0.053	0.011	0.103	0.052	0.010
	401	0.099	0.051	0.010	0.106	0.053	0.011	0.100	0.050	0.010
	801	0.103	0.052	0.011	0.101	0.050	0.010	0.099	0.051	0.011
	1601	0.099	0.050	0.010	0.102	0.051	0.010	0.101	0.051	0.010
0.08	101	0.107	0.056	0.012	0.112	0.058	0.012	0.106	0.052	0.010
	201	0.102	0.052	0.011	0.106	0.053	0.011	0.105	0.053	0.011
	401	0.102	0.051	0.011	0.105	0.054	0.011	0.102	0.052	0.010
	801	0.104	0.053	0.010	0.102	0.051	0.009	0.103	0.052	0.010
	1601	0.102	0.052	0.011	0.106	0.053	0.011	0.103	0.052	0.010
0.10	101	0.107	0.054	0.011	0.118	0.059	0.013	0.108	0.055	0.011
	201	0.104	0.053	0.011	0.109	0.055	0.011	0.104	0.052	0.010
	401	0.104	0.052	0.011	0.110	0.056	0.012	0.105	0.053	0.011
	801	0.103	0.050	0.011	0.105	0.053	0.010	0.102	0.050	0.010
	1601	0.100	0.050	0.010	0.101	0.052	0.010	0.100	0.050	0.011

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and α = theoretical false positive rate.

Table B3. *Empirical false positive rates for Model 3: (REPCFA, $K = 133$)*

		Modification Subset 1 ($k = 1$)			**Modification Subset 2 ($k = 9$)**			**Modification Subset 3 ($k = 21$)**		
RMSEA	**Sample Size**	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.00	101	0.093	0.045	0.008	0.085	0.042	0.008	0.079	0.037	0.007
	201	0.097	0.048	0.010	0.095	0.046	0.009	0.092	0.046	0.009
	401	0.097	0.051	0.009	0.096	0.046	0.009	0.095	0.046	0.010
	801	0.101	0.051	0.010	0.098	0.049	0.010	0.097	0.049	0.009
	1601	0.101	0.051	0.010	0.099	0.049	0.009	0.098	0.048	0.010
0.03	101	0.091	0.045	0.009	0.087	0.042	0.007	0.083	0.040	0.007
	201	0.100	0.051	0.011	0.095	0.047	0.010	0.093	0.046	0.010
	401	0.098	0.049	0.010	0.098	0.049	0.009	0.099	0.051	0.010
	801	0.098	0.050	0.010	0.099	0.047	0.009	0.097	0.048	0.010
	1601	0.099	0.050	0.009	0.100	0.050	0.010	0.100	0.051	0.011
0.05	101	0.099	0.048	0.009	0.092	0.045	0.009	0.088	0.043	0.008
	201	0.100	0.049	0.010	0.096	0.048	0.010	0.095	0.047	0.010
	401	0.097	0.049	0.010	0.101	0.051	0.010	0.095	0.047	0.009
	801	0.102	0.051	0.011	0.099	0.049	0.011	0.099	0.049	0.009
	1601	0.099	0.050	0.011	0.099	0.049	0.009	0.101	0.051	0.010
0.08	101	0.096	0.047	0.009	0.097	0.048	0.009	0.092	0.045	0.009
	201	0.097	0.049	0.009	0.103	0.053	0.010	0.099	0.050	0.010
	401	0.103	0.052	0.010	0.099	0.051	0.010	0.100	0.050	0.010
	801	0.101	0.050	0.010	0.104	0.052	0.010	0.102	0.052	0.010
	1601	0.102	0.052	0.010	0.100	0.048	0.009	0.101	0.051	0.010
0.10	101	0.099	0.048	0.010	0.097	0.046	0.009	0.097	0.047	0.009
	201	0.099	0.049	0.009	0.104	0.052	0.010	0.100	0.049	0.010
	401	0.101	0.051	0.010	0.102	0.053	0.011	0.103	0.051	0.011
	801	0.099	0.050	0.010	0.100	0.050	0.010	0.101	0.051	0.010
	1601	0.101	0.050	0.010	0.101	0.051	0.010	0.104	0.053	0.011

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and α = theoretical false positive rate.

Appendix C. Asymptotic Power Analysis Results

Table C1. *Approximation error proportion (π_{diff}) for Model 1: (CFA, $K = 25$)*

		Modification Subset 1 ($k = 1$)		Modification Subset 2 ($k = 3$)		Modification Subset 3 ($k = 6$)	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	-	-	-	-	-	-
	201	0.93	-	-	-	-	-
	401	0.54	0.96	0.83	-	-	-
	801	0.35	0.58	0.56	0.82	0.76	-
	1601	0.25	0.39	0.43	0.59	0.61	0.77
0.05	101	0.71	-	-	-	-	-
	201	0.43	0.74	0.68	-	0.90	-
	401	0.29	0.47	0.49	0.69	0.67	0.88
	801	0.22	0.33	0.39	0.52	0.56	0.70
	1601	0.19	0.26	0.34	0.43	0.51	0.60
0.08	101	0.37	0.62	0.59	0.88	0.80	-
	201	0.26	0.41	0.44	0.62	0.62	0.80
	401	0.21	0.30	0.37	0.48	0.54	0.65
	801	0.18	0.24	0.33	0.40	0.50	0.57
	1601	0.16	0.21	0.31	0.36	0.47	0.53
0.10	101	0.29	0.47	0.48	0.69	0.67	0.88
	201	0.22	0.33	0.39	0.52	0.56	0.69
	401	0.19	0.26	0.34	0.43	0.51	0.60
	801	0.17	0.22	0.32	0.38	0.48	0.54
	1601	0.16	0.19	0.31	0.35	0.47	0.51

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

Table C2. *Approximation error proportion (π_{diff}) for Model 2: (LVPM, $K = 86$)*

		Modification Subset 1 ($k = 1$)		**Modification Subset 2 ($k = 6$)**		**Modification Subset 3 ($k = 15$)**	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	0.52	-	-	-	-	-
	201	0.29	0.55	0.59	0.96	0.87	-
	401	0.17	0.30	0.37	0.57	0.57	0.81
	801	0.10	0.18	0.25	0.37	0.43	0.56
	1601	0.07	0.12	0.20	0.27	0.35	0.43
0.05	101	0.22	0.41	0.46	0.73	0.70	-
	201	0.13	0.24	0.30	0.45	0.49	0.67
	401	0.09	0.15	0.22	0.31	0.38	0.49
	801	0.07	0.10	0.18	0.24	0.33	0.40
	1601	0.06	0.08	0.16	0.20	0.30	0.35
0.08	101	0.11	0.20	0.27	0.39	0.44	0.59
	201	0.08	0.13	0.21	0.28	0.36	0.45
	401	0.06	0.09	0.17	0.22	0.32	0.38
	801	0.05	0.07	0.16	0.19	0.30	0.34
	1601	0.05	0.06	0.15	0.17	0.29	0.31
0.10	101	0.09	0.15	0.22	0.31	0.38	0.49
	201	0.07	0.10	0.18	0.24	0.33	0.40
	401	0.06	0.08	0.16	0.20	0.30	0.35
	801	0.05	0.06	0.15	0.18	0.29	0.32
	1601	0.05	0.06	0.15	0.16	0.28	0.30

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

Table C3. *Approximation error proportion (π_{diff}) for Model 3: (REPCFA, $K = 133$)*

		Modification Subset 1 ($k = 1$)		**Modification Subset 2 ($k = 9$)**		**Modification Subset 3 ($k = 21$)**	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	0.34	0.67	0.81	-	-	-
	201	0.19	0.36	0.47	0.74	0.69	-
	401	0.11	0.20	0.30	0.45	0.46	0.64
	801	0.07	0.12	0.21	0.30	0.35	0.45
	1601	0.05	0.08	0.17	0.22	0.29	0.36
0.05	101	0.14	0.27	0.37	0.57	0.56	0.80
	201	0.09	0.15	0.25	0.36	0.40	0.54
	401	0.06	0.10	0.19	0.25	0.32	0.40
	801	0.04	0.07	0.16	0.20	0.28	0.33
	1601	0.04	0.05	0.14	0.17	0.26	0.29
0.08	101	0.07	0.13	0.22	0.31	0.36	0.48
	201	0.05	0.08	0.17	0.23	0.30	0.37
	401	0.04	0.06	0.15	0.18	0.27	0.31
	801	0.03	0.05	0.14	0.16	0.25	0.28
	1601	0.03	0.04	0.13	0.15	0.24	0.26
0.10	101	0.06	0.10	0.19	0.25	0.32	0.40
	201	0.04	0.07	0.16	0.20	0.28	0.33
	401	0.04	0.05	0.14	0.17	0.26	0.29
	801	0.03	0.04	0.13	0.15	0.25	0.27
	1601	0.03	0.04	0.13	0.14	0.24	0.26

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

Appendix D. Simulation Study 2 Results

Table D1. *Empirical power for Model 1: (CFA, $K = 25$)*

		Modification Subset 1 ($k = 1$)		**Modification Subset 2 ($k = 3$)**		**Modification Subset 3 ($k = 6$)**	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	-	-	-	-	-	-
	201	0.49	-	-	-	-	-
	401	0.50	0.80	0.49	-	-	-
	801	0.50	0.80	0.50	0.80	0.50	-
	1601	0.50	0.80	0.49	0.79	0.50	0.80
0.05	101	0.48	-	-	-	-	-
	201	0.49	0.79	0.50	-	0.50	-
	401	0.50	0.80	0.50	0.79	0.50	0.79
	801	0.50	0.79	0.49	0.79	0.50	0.79
	1601	0.50	0.80	0.50	0.81	0.50	0.80
0.08	101	0.50	0.79	0.48	0.78	0.49	-
	201	0.50	0.79	0.50	0.79	0.50	0.78
	401	0.50	0.80	0.50	0.80	0.50	0.80
	801	0.50	0.80	0.49	0.79	0.51	0.79
	1601	0.50	0.80	0.50	0.80	0.49	0.81
0.10	101	0.51	0.78	0.49	0.78	0.49	0.79
	201	0.50	0.79	0.50	0.79	0.50	0.80
	401	0.50	0.80	0.50	0.79	0.50	0.79
	801	0.49	0.80	0.49	0.79	0.50	0.80
	1601	0.50	0.79	0.51	0.80	0.50	0.79

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

Table D2. *Empirical power for Model 2: (LVPM, $K = 86$)*

		Modification Subset 1 ($k = 1$)		**Modification Subset 2 ($k = 6$)**		**Modification Subset 3 ($k = 9$)**	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	0.48	-	-	-	-	-
	201	0.50	0.79	0.49	0.79	0.47	-
	401	0.50	0.80	0.50	0.79	0.50	0.80
	801	0.50	0.80	0.50	0.80	0.49	0.80
	1601	0.50	0.80	0.50	0.80	0.50	0.80
0.05	101	0.48	0.78	0.48	0.77	0.47	-
	201	0.48	0.79	0.49	0.79	0.49	0.79
	401	0.50	0.80	0.51	0.80	0.50	0.79
	801	0.49	0.80	0.51	0.80	0.50	0.80
	1601	0.50	0.80	0.51	0.80	0.51	0.80
0.08	101	0.49	0.77	0.48	0.78	0.49	0.77
	201	0.50	0.79	0.50	0.80	0.50	0.79
	401	0.50	0.79	0.51	0.80	0.50	0.80
	801	0.51	0.80	0.50	0.81	0.51	0.80
	1601	0.49	0.80	0.50	0.79	0.51	0.81
0.10	101	0.49	0.78	0.49	0.79	0.49	0.78
	201	0.49	0.79	0.51	0.80	0.50	0.80
	401	0.49	0.79	0.51	0.79	0.51	0.80
	801	0.50	0.79	0.51	0.81	0.51	0.80
	1601	0.50	0.80	0.51	0.80	0.51	0.80

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

Table D3. *Empirical power for Model 3: (REPCFA, $K = 133$)*

		Modification Subset 1 ($k = 1$)		**Modification Subset 2 ($k = 9$)**		**Modification Subset 3 ($k = 21$)**	
RMSEA	**Sample Size**	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$	$1 - \beta = 0.5$	$1 - \beta = 0.8$
0.03	101	0.47	0.78	0.44	-	-	-
	201	0.49	0.79	0.47	0.78	0.47	-
	401	0.49	0.80	0.49	0.79	0.48	0.79
	801	0.49	0.79	0.49	0.80	0.50	0.80
	1601	0.49	0.80	0.49	0.79	0.50	0.80
0.05	101	0.47	0.78	0.46	0.76	0.42	0.74
	201	0.48	0.79	0.48	0.78	0.48	0.78
	401	0.49	0.79	0.50	0.79	0.49	0.78
	801	0.49	0.80	0.50	0.80	0.50	0.80
	1601	0.50	0.81	0.50	0.80	0.50	0.80
0.08	101	0.47	0.78	0.46	0.77	0.45	0.76
	201	0.49	0.80	0.49	0.79	0.49	0.78
	401	0.50	0.80	0.50	0.80	0.50	0.79
	801	0.51	0.80	0.51	0.80	0.51	0.80
	1601	0.50	0.80	0.50	0.80	0.50	0.81
0.10	101	0.48	0.78	0.47	0.78	0.47	0.76
	201	0.50	0.79	0.50	0.80	0.49	0.79
	401	0.50	0.80	0.50	0.80	0.49	0.79
	801	0.50	0.80	0.51	0.80	0.51	0.81
	1601	0.50	0.80	0.51	0.80	0.50	0.80

Note: K = model degrees of freedom, k = modification subset degrees of freedom, and $1 - \beta$ = theoretical power, - denotes missing values due to test being underpowered.

References

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. https://doi.org/10.1177/0049124192020003002

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

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

Bentler, P. M. (1995). *EQS Structural Equations Program Manual*. Multivariate Software.

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

Bollen, K. A. (1989). *Structural Equations with Latent Variables* (1st ed.). Wiley. https://doi.org/10.1002/9781118619179

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics* (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

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

Breivik, E., & Olsson, U. H. (2001). Adding variables to improve model fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 169–194). Scientific Software International.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421. https://doi.org/10.1037/1082-989X.7.4.403

Chen, J., Variyath, A. M., & Abraham, B. (2008). Adjusted Empirical Likelihood and its Properties. *Journal of Computational and Graphical Statistics*, *17*(2), 426–443. https://doi.org/10.1198/106186008X321068

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Fang, K.-T., & Zhang, Y.-T. (1990). *Generalized multivariate analysis*. Science Press.

Fornell, C., & Yi, Y. (1992). Assumptions of the Two-Step Approach to Latent Variable Modeling. *Sociological Methods & Research*, *20*(3), 291–320. https://doi.org/10.1177/0049124192020003001

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-BA117A

Genz, A., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Springer-Verlag.

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. https://doi.org/10.1177/0013164410384856

Harring, J. R., Weiss, B. A., & Hsu, J.-C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods*, *17*(2), 193–214. https://doi.org/10.1037/a0027539

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), 319–336. https://doi.org/10.1037/a0024917

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

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage Publications, Inc.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362. https://doi.org/10.1037/0033-2909.112.2.351

Hult, H., & Lindskog, F. (2002). Multivariate Extremes, Aggregation and Dependence in Elliptical Distributions. *Advances in Applied Probability*, *34*(3), 587–608.

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

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.80. *Lincolnwood, IL: Scientific Software International Inc*.

Kariya, T., & Eaton, M. L. (1977). Robust Tests for Spherical Symmetry. *The Annals of Statistics*, *5*(1), 206–215. JSTOR.

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. https://doi.org/10.1177/0049124114543236

Kenny, D. A., & McCoach, D. B. (2003). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 333–351. https://doi.org/10.1207/S15328007SEM1003_1

Klein, A., & Moosbrugger, H. (2000). Maximum Likelihood Estimation of Latent Interaction Effects with the LMS Method. *Psychometrika*, *65*(4), 457–474. https://doi.org/10.1007/BF02296338

King, M. L. (1980). Robust Tests for Spherical Symmetry and Their Application to Least Squares Regression. *The Annals of Statistics*, *8*(6). https://doi.org/10.1214/aos/1176345199

Lai, K. (2019). Creating Misspecified Models in Moment Structure Analysis. *Psychometrika*, *84*, 781–801. https://doi.org/10.1007/s11336-018-09655-0

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*, 220–239.

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), 388–404. https://doi.org/10.1037/met0000068

Lee, S.-Y., & Zhu, H.-T. (2002). Maximum Likelihood Estimation of Nonlinear Structural Equation Models. *Psychometrika*, *67*(2), 189–210. https://doi.org/10.1007/BF02294842

Lenth, R. V. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Luijben, T. C. W. (1991). Equivalent models in covariance structure analysis. *Psychometrika*, *56*(4), 653–665. https://doi.org/10.1007/BF02294497

MacCallum, R. (1986). Specification Searches in Covariance Structure Modeling. *Psychological Bulletin*, *100*(1), 107–120.

MacCallum, R. C. (2003). 2001 Presidential Address: Working with Imperfect Models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

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

Maydeu-Olivares, A. (2017). Assessing the Size of Model Misfit in Structural Equation Models. *Psychometrika*, *82*, 533–558. https://doi.org/10.1007/s11336-016-9552-7

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. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Moshagen, M. (2012). The Model Size Effect in SEM: Inflated Goodness-of-Fit Statistics Are Due to the Size of the Covariance Matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Mueller, R. O., & Hancock, G. R. (2008). Best Practices in Structural Equation Modeling. In J. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 488–508). SAGE Publications, Inc. https://doi.org/10.4135/9781412995627.d38

Muthén, B. O. (2002). Beyond SEM: General Latent Variable Modeling. *Behaviormetrika*, *29*(1), 81–117. https://doi.org/10.2333/bhmk.29.81

Orsi, C. (2017). *New insights into non-central beta distributions* (No. arXiv:1706.08557). arXiv. http://arxiv.org/abs/1706.08557

Preacher, K. J. (2006). Quantifying Parsimony in Structural Equation Modeling. *Multivariate Behavioral Research*, *41*(3), 227–259. https://doi.org/10.1207/s15327906mbr4103_1

Preacher, K. J., & Haley, Y. E. (2023). Model selection in structural equation modeling. In *Handbook of structural equation modeling* (pp. 206–222).

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. Wiley.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rosseel, Y., & Loh, W. W. (2024). A structural after measurement approach to structural equation modeling. *Psychological methods*, *29*(3), 561–588. https://doi.org/10.1037/met0000503

Saris, W. E., Satorra, A., & Sorbom, D. (1987). The Detection and Correction of Specification Errors in Structural Equation Models. *Sociological Methodology*, *17*, 105. https://doi.org/10.2307/271030

Saris, W. E., Satorra, A., & Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561-582,. https://doi.org/10.1080/10705510903203433

Satorra, A. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *American statistical association 1988 proceedings of business and economics sections* (pp. 308-313). American Statistical Association.

Satorra, A. (2015). A Comment on a Paper by H. Wu and M. W. Browne (2014). *Psychometrika*, *80*(3), 613–618. https://doi.org/10.1007/s11336-015-9455-z

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 339–419). Sage Publications, Inc.

Savalei, V. (2012). The Relationship Between Root Mean Square Error of Approximation and Model Misspecification in Confirmatory Factor Analysis Models. *Educational and Psychological Measurement*, *72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*(2), 150–170. https://doi.org/10.1037/1082-989X.13.2.150

Shapiro, A. (1986). Asymptotic Theory of Overparameterized Structural Models. *Journal of the American Statistical Association*, *81*(393), 142–149.

Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2022). Evaluating SEM Model Fit with Small Degrees of Freedom. *Multivariate Behavioral Research*, *57*(2–3), 179–207. https://doi.org/10.1080/00273171.2020.1868965

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement*, *79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Sörbom, D. (1974). A GENERAL METHOD FOR STUDYING DIFFERENCES IN FACTOR MEANS AND FACTOR STRUCTURE BETWEEN GROUPS. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

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

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 777–781. https://doi.org/10.1080/10705511.2016.1217487

Steiger, J. H., & Lind, J. C. (1980). Statistically Based Tests for the Number of Common Factors. Paper Presented at the Psychometric Society Annual Meeting, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, *50*(3), 253–263. https://doi.org/10.1007/BF02294104

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika*, *66*(4), 473–485. https://doi.org/10.1007/BF02296190

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, *23*(1), 27–41. https://doi.org/10.1037/met0000147

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459. https://doi.org/10.1007/BF02290601

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://doi.org/10.1007/BF02291170

Tukey, J. W. (1961). Discussion, Emphasizing the Connection Between Analysis of Variance and Spectrum Analysis. *Technometrics*, *3*(2), 191–219. https://doi.org/10.1080/00401706.1961.10489940

Van De Schoot, R., Hoijtink, H., & Dekovic, M. (2010). Testing Inequality Constrained Hypotheses in SEM Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 443–463. https://doi.org/10.1080/10705511.2010.489010

West, S. G., Wu, W., McNeish, D., & Savord, A. (2023). Model fit in structural equation modeling. In *Handbook of structural equation modeling* (pp. 184–205).

Whittaker, T. A. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *The Journal of Experimental Education*, *80*(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wu, H., & Browne, M. W. (2015). Quantifying Adventitious Error in a Covariance Structure as a Random Effect. *Psychometrika*, *80*(3), 571–600. https://doi.org/10.1007/s11336-015-9451-3

Yuan, K.-H. (2005). Fit Indices Versus Test Statistics. *Multivariate Behavioral Research*, *40*(1), 115–148. https://doi.org/10.1207/s15327906mbr4001_5

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309. https://doi.org/10.1111/j.2044-8317.1998.tb00682.x

Yuan, K.-H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*(3), 225–243.

Yuan, K.-H., & Bentler, P. M. (2000). 5. Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yuan, K.-H., & Bentler, P. M. (2006). Structural Equation Modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 297–358). Elsevier. https://doi.org/10.1016/S0169-7161(06)26010-3

Yuan, K.-H., Hayashi, K., & Bentler, P. M. (2007). Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses. *Journal of Multivariate Analysis*, *98*(6), 1262–1282. https://doi.org/10.1016/j.jmva.2006.08.005

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical Correction to the Likelihood Ratio Statistic for Structural Equation Modeling with Many Variables. *Psychometrika*, *80*(2), 379–405. https://doi.org/10.1007/s11336-013-9386-5