The Unified Approach for Model Evaluation in Structural Equation Modeling

By

Sunthud Pornprasertmanit

Submitted to the Department of Psychology and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Wei Wu, Chairperson

Todd D. Little

Carol Woods

Pascal Deboeck

William Skorupski

Date defended: May 16, 2014

The Dissertation Committee for Sunthud Pornprasertmanit certifies
that this is the approved version of the following dissertation :

The Unified Approach for Model Evaluation in Structural Equation Modeling

Wei Wu, Chairperson

Date approved: May 16, 2014

Abstract

Practical fit indices have been widely used for model fit evaluation in Structural Equation Modeling. This dissertation discusses the properties of the fit indices including their influencing factors. These properties prevent researchers from deriving one-size-fit-all cutoffs for the fit indices. In addition, the past simulation studies on model fit evaluation have several limitations. The major limitation is that most studies have focused on test of exact fit rather than approximate fit which is not consistent with the goal of practical fit indices. This dissertation reviews alternative approaches to account for the limitations and proposes a unified method for model fit evaluation combining the advantages of the alternative approaches. The unified approach allows researchers to test approximate fit and take into account sampling error in model evaluation.

Two simulation studies are conducted to investigate the performance of the unified approach comparing to the other model fit evaluation methods. Two types of models are included in this study: confirmatory factor analysis and growth curve models. The results show that the unified approach appropriately rejects severely misspecified models and retains trivially misspecified models across all types of misspecification. Furthermore, the rejection rates are negligibly influenced by model characteristics and sample size. The other model evaluation methods do not have all of the desired properties described above. The unified approach, however, does not always provide model decision when sample size is low or when the level of maximal trivial misspecification specified by users is close to the actual degree of misspecification. If sample size is high and the level of specified maximal trivial misspecification is either lower or higher than the actual degree of misspecification, the unified approach is able to decide be-

tween model retention and model rejection. The extensions of the unified approach for nonnormal distribution, missing data, or nested model comparison are provided.

Keywords: structural equation modeling, model parsimony, model fit, practical fit indices

Acknowledgements

I would like to express my appreciation to my advisor, Wei Wu, who has supported me to accomplish this challenging project. She knew how to support me. The idea behind this dissertation cannot be completed if she did not help me polish my ideas. I appreciate her time reading and providing great comments for the draft. I am honored to have Wei Wu as my advisor.

I would like to express my appreciation to my committee members, Todd Little, William Skorupski, Carol Woods, and Pascal DeBoeck, for their help and valuable comments. They have built a great quantitative program that I am very honored to be part of.

Thanks to Adam Hafdahl for introducing me to the theory of equivalence testing that extremely improved my idea of the unified approach.

Thanks to KU Center for Research Methods and Data Analysis for the computational resources and the opportunity to work at the center. The experience at the center allowed me to learn and invent research ideas.

Thanks to all friends and colleagues in KU for their selfless supports.

Special thanks to my family for the continuing support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Structural equation modeling (SEM) has been one of the most popular frameworks to represent theoretical relations among observed and unobserved variables, such as factor structures underlying multiple items, predicting relationships among variables or factors, or growth trajectories. The models are then evaluated regarding whether they provide a plausible explanation of the relationships among the variables, which is referred to as model fit evaluation. The accuracy of model fit evaluation would highly influence the decision of retaining a hypothesized model. Unfortunately, since SEM was developed, no consensus has been reached in terms of the best way to evaluate model fit (Lance et al., 2006). The current practice of model fit evaluation also has many problems. This dissertation aims to discuss these problems and propose ways to improve the current practices of model evaluation.

The organization of the proposal as follows. In this chapter, I start with a brief discussion of the history of model evaluation in SEM. I then review the properties of practical fit indices that are currently used in model evaluation and the limitations of the fit indices including the problems of using cutoffs. In Chapter 2, I review alternative methods that account for most of the limitations discussed in Chapter 1. In Chapter 3, a unified approach for model evaluation is proposed. I also discuss how approximate fit and severe fit are defined. The definitions of the degrees of model fit are required by the alternative methods and the unified approach. An empirical

example of applying the unified approach is provided. In Chapter 4, I propose two simulation studies for investigating the performance of the unified approach and comparing it with the other model evaluation approaches. In Chapter 5, the simulation results are provided. In Chapter 6, I discuss whether the unified approach solve the problems of the use of fit indices and the alternative methods. I also discuss the limitations of the unified approach and the simulation studies in this dissertation. The extensions of the unified approach for missing data, nonnormal data, and nested model comparison are also provided.

1.1 A Short History

The history of model fit evaluation in SEM is provided in order to understand why the current method of model evaluation is used and needs to be improved. The initial test statistic for model fit evaluation of SEMs is a chi-square test statistic. The chi-square statistic is a function of the discrepancy between model-implied sufficient statistics (mean vector and covariance matrix computed from parameter estimates) and sample sufficient statistics (sample mean vector and covariance matrix). When a hypothesized model is the same as a population model, the chi-square statistic follows a central chi-square distribution. In this case, the difference in chi-square statistic is purely due to sampling error. If the chi-square statistic is not significant, a hypothesized model is preferred.

When a hypothesized model is a population model, the chi-square statistic leads to accurate statistical inference regardless of sample size (Bollen, 1989; Maydeu-Olivares & Cai, 2006; McIntosh, 2007). That is, the type I error rate (i.e., the proportion of rejecting a hypothesized model if the model is in fact a population model) is close to nominal level (e.g., .05). However, if the hypothesized model is trivially misspecified (e.g., a population having a small cross loading but not specified in a hypothesized model), the chi-square statistic will result in rejection of the hypothesized model when sample size is large enough (Bentler, 2007; Maydeu-Olivares & Cai, 2006). This characteristic is not desirable because researchers wish to retain a hypothesized model that

approximates a population well even if it is not exactly equal to the population model (MacCallum, 2003; MacCallum & Austin, 2000). Thus, alternative measures for model fit evaluation have been developed, such as the ratio of chi-square statistic to degree of freedom (Wheaton et al., 1977), goodness-of-fit index (GFI; Jöreskog & Sörbom, 1981), normed fit index (NFI; Bentler & Bonett, 1980), and so on. These alternative measures are often referred to as practical fit indices.

More than 30 practical fit indices have been developed in the past decades (Marsh et al., 1988). The primary goal of the practical fit indices is to quantify the amount of misfit between model-impled and sample sufficient statistics that is not sensitive to sample size. In fact, many of the practical fit indices, including root mean square error of approximation (RMSEA; Browne & Cudeck, 1992; Steiger & Lind, 1980), comparative fit index (CFI; Bentler, 1990), and Tucker-Lewis index (TLI; Tucker & Lewis, 1973), are not sensitive to sample size, except for small sample size (Hu & Bentler, 1998). Although the fit indices were developed to represent the degree of misfit in a continuous scale, cutoff criteria are proposed for the fit indices to facilitate the decision of whether a hypothesized model sufficiently fits the data based on which researchers will know whether to retain or revise the hypothesized model (Barrett, 2007).

Although many guidelines are available for the cutoffs of the fit indices, they were mostly established based on personal experience (Bentler & Bonett, 1980; Browne & Cudeck, 1992). The suggested cutoffs from these guidelines have been frequently used in practice (Lance et al., 2006) even though researchers do not really know what Type I error rates and Type II error rates are from using the cutoffs. Type I error rate is the proportion of falsely rejecting correct models. Type II error rate is the proportion of falsely accepting severely misspecified models. Thus, Hu & Bentler (1999) conducted a large scale simulation study to search for the optimal cutoffs for some of the fit indices that would minimize both Type I and II errors. The examples of the proposed cutoffs are RMSEA of .06 and CFI of .95. They also proposed a two-index strategy to combine information from different fit indices together.

Hu and Bentler's simulation was not perfect. The proposed cutoffs were not applicable to other models beside the models used in Hu and Bentler's simulation, such as confirmatory factor

analysis (CFA) models with a large indicators to factors ratio (Heene et al., 2011), latent growth model (Wu & West, 2010), and CFA models with categorical variables (Nye & Drasgow, 2011). The proposed cutoffs may lead to inflated Type I error or inflated Type II error. Because of this problem, Barrett (2007) recommended that researchers abandon all practical fit indices and use only the chi-square test statistic for model fit evaluation (with care). Although researchers agreed that the use of practical fit indices cutoffs is not legitimate, they criticized Barrett's ideas and argued that fit indices are still useful in many aspects (Bentler, 2007; McIntosh, 2007; Millsap, 2007; Mulaik, 2007; Steiger, 2007). In this chapter, I will provide a thorough overview of the benefits and the limitations of fit indices.

In sum, no clear cutoff has been established for the practical fit indices. There is no way for researchers to know the Type I or II errors associated with the cutoffs. On the other hand, although the chi-square statistic has a clear cutoff, it is sensitive to trivial misspecification with large sample size. Thus practical fit indices are still very useful in compensating the problem associated with the chi-square test statistic. This paper will propose a unified approach that borrows strengths from both chi-square and practical fit indices to improve the use of practical fit indices.

1.2 Properties of Practical Fit Indices

In this section, the definition of practical fit indices and their properties are reviewed. Some properties are desirable so practical fit indices should continue to be used. Other properties are undesirable so users need to be careful in using them. Moreover, the discussion of the properties is crucial to design an appropriate use of fit indices in model evaluation.

1.2.1 Effect Size Measures

One of the most important characteristics of fit indices is that they are effect size measures. Fit indices are the quantitative measures of the amount of misfit between a hypothesized model and sample data (sample fit indices) or the population model underlying the sample data (population

fit indices). Similar to other effect size statistics (e.g., effect size measures for two-group mean differences), there are multiple ways to quantify the degrees of misfit. In this section, I will use the population definition of all practical fit indices because the population formula provides direct interpretation of the meanings behind each fit indices (excluding the correction of biased estimators).

Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the population mean vector and covariance matrix and $\boldsymbol{\mu}_M$ and $\boldsymbol{\Sigma}_M$ be the model-implied means and covariance matrix after fitting $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to a hypothesized model M. The discrepancy (F_M) between the two means and covariance matrices can be computed using Equation 1.1 (Kenny & McCoach, 2003):

$$F_M = \text{tr}\left\{\boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}_M\right]^{-1}\right\} - \log\left|\boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}_M\right]^{-1}\right| - N + \left[\boldsymbol{\mu} - \boldsymbol{\mu}_M\right]'\left[\boldsymbol{\Sigma}_M\right]^{-1}\left[\boldsymbol{\mu} - \boldsymbol{\mu}_M\right], \tag{1.1}$$

where N is sample size. F_M can be calculated by any SEM packages by using population mean vector and covariance matrix as data to fit a SEM model. The resulting chi-square value is the noncentrality parameter (λ_M). The relationship between the noncentrality parameter and the discrepancy function is defined by Equation 1.2 (Browne & Cudeck, 1992; Satorra & Saris, 1985; Saris & Satorra, 1993):

$$\lambda_M = (N-1)F_M. \tag{1.2}$$

That is, the discrepancy value can be calculated by dividing the chi-square value by sample size minus 1.

Most fit indices are defined as a function of the discrepancy value. Fit indices are classified into two groups: absolute fit and relative (incremental) fit. Absolute fit is a class of fit indices that is defined by only statistics from fitting a hypothesized model. Two popular absolute fit indices are RMSEA and standardized root mean square residual (SRMR). RMSEA is defined as the amount of population misfit per degree of freedom. The population value of RMSEA for Model M (ε_M) can be computed as follows:

$$\varepsilon_M = \sqrt{\frac{F_M}{df_M}}, \tag{1.3}$$

where df_M is the degree of freedom of Model M. SRMR is computed as follows:

$$SRMR = \sqrt{\sum_j \sum_{k \leq j} \frac{\left(\frac{\sigma_{Mjk}}{\sqrt{\sigma_{Mjj}}\sqrt{\sigma_{Mkk}}} - \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}} \right)^2}{\frac{p(p+1)}{2}}}, \tag{1.4}$$

where σ_{Mjk} and σ_{jk} are the j^{th}-row and k^{th}-column element of $\boldsymbol{\Sigma}_M$ and $\boldsymbol{\Sigma}$, respectively, and p is the number of variables. SRMR can be interpreted as the average difference between the population correlation matrix and model-implied correlation matrix. As shown in Browne et al. (2002), RMSEA is a weighted average of correlation residuals for models without mean structure. In comparison, SRMR is an unweighted average across the correlation elements. Both indices are effect size measures that provide different information about the misfits of a hypothesized model.

Relative fit indices quantify misfit of a hypothesized model by comparing it to a worst-fit model (or a baseline model). Most relative fit indices are the estimates of two population values (Mahler, 2011). The first population value, often referred to as population CFI, is the ratio of two differences in discrepancy values:

$$CFI = \frac{F_0 - F_M}{F_0 - F_S} = 1 - \frac{F_M}{F_0}, \tag{1.5}$$

where F_0, F_M, and F_S are the discrepancy values of a baseline model, a hypothesized model, and a saturated model, respectively. The discrepancy value of the saturated model is 0 because $\boldsymbol{\mu} = \boldsymbol{\mu}_M$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_M$. Thus, the formula can be expressed as the latter term in Equation 1.5. CFI is then interpreted as the degree of close fit from a hypothesized model relative to a baseline model. The sample values of CFI (Bentler, 1990), NFI (Bentler & Bonett, 1980), Relative Noncentrality Index (RNI; McDonald & Marsh, 1990), and Incremental Fit Index (IFI; Bollen, 1989) are all estimates of this population value.

The second population value is the population TLI, which is CFI corrected for model complex-

ity. Here model complexity is usually indicated by model degrees of freedom:

$$TLI = \frac{F_0/df_0 - F_M/df_M}{F_0/df_0} = 1 - \frac{F_M}{F_0} \cdot \frac{df_0}{df_M}, \tag{1.6}$$

where df_M and df_0 are degrees of freedom of a hypothesized model and a baseline model, respectively. The term df_M/df_0 is referred to as a parsimony ratio (Mulaik et al., 1989). A model with more constraints will have a larger parsimony ratio. Population TLI will indicate better fit for models with higher parsimony ratio holding discrepancy value constant. Population TLI is then interpreted as the degree of close fit from a hypothesized model relative to a baseline model accounting for model parsimony. The sample values of TLI (Tucker & Lewis, 1973) and Relative Fit Index (RFI; Bollen, 1986) are both estimates of this population value.

Both types of population relative fit indices require the definition of a baseline model. The baseline model is usually the model that estimates only means and variances (covariances are fixed as 0). Note that there are alternative baseline models (see Rigdon, 1996; Widaman & Thompson, 2003; Williams & O'Boyle, 2011), as well as alternative saturated models (Williams & O'Boyle, 2011).

Although the guidelines for interpreting the magnitude of effect sizes are available (e.g., RMSEA by Browne & Cudeck, 1992), the guidelines are dependent on research context (Cohen, 1988, 1992). For example, if the outcome is death or serious issues, a trivial effect size (e.g., R^2 of .01) can be considered as nontrivial (Ferguson, 2009). Similar to any effect size measures, a certain value of a fit index can be considered as an indication of severe misfit in some contexts but trivial in the others. For example, the model predicting the death outcome is needed to be more accurate than others. Thus, researchers should not use any guidelines for the interpretation of any effect size measures unless they find those guidelines appropriate for their research contexts.

1.2.2 Influencing Factors

In this section, the factors influencing practical fit indices are reviewed. Three types of model characteristics (i.e., model size, type of models, and incidental parameters) can influence the population values of a fit index. These factors will influence practical fit indices regardless of the estimation method. On the other hand, sample size, estimation method, and data distribution would influence practical fit indices through estimation. These factors may influence the expected value of the fit indices or the shape of their sampling distributions (e.g., variances).

1.2.2.1 Model Size

In the previous studies, model size has been represented by degree of freedom, the number of observed variables, the number of elements in covariance matrix, or the number of indicators per factor. As degree of freedom increases, RMSEA, CFI, and TLI tend to decrease. Holding the other model characteristics constant, the number of observed variables also tend to decrease RMSEA (better fit), CFI, and TLI (poorer fit; Ding et al., 1995; Heene et al., 2011; Kenny & McCoach, 2003; Savalei, 2012). The reason is that most of the past simulation studies used independent-clustered CFA models in which each variable is loaded on only one factor. Thus, when the number of observed variables increases, the number of constrained cross-loadings and correlated errors also increases, leading to an increase in degree of freedom. Model size has a positive relationship with the number of free parameters in general so fit indices may be influenced by the number of free parameters. Moshagen (2012), however, showed that the size of a covariance matrix (model size), instead of the number of free parameters, was related to the change in fit indices .

1.2.2.2 Model Type

Most simulation studies used confirmatory factor analysis (CFA) to derive fit indices cutoffs. CFA is only one of the simplest models in SEM. In practice, SEM can be much more complicated. For example, growth curve models involved mean structure in addition to covariance structure. Wu (2008) and Wu & West (2010) showed that different fit indices had differential performance in

detecting different types of misspecification in growth curve models. This problem applies to other longitudinal models, such as autoregressive model (Liu et al., 2012; Raykov, 2000). As another example, multilevel SEMs involve misfits at different levels of data structure (Ryu & West, 2009). The distributions of fit indices were narrower and closer to the point of perfect fit at the macro level than those at the micro level. As a result, the same criteria for model evaluation from the micro level are not applicable for the macro level. Otherwise, misspecified models at the macro level are retained at a higher rate than misspecified models at the micro level (Boulton, 2011; Ryu & West, 2009).

1.2.2.3 Incidental Parameters

Given the same model type, the parameter values in the model, referred to as incidental parameters, affect the amount of misfit due to fixed parameters. That is, the same amount of misspecified parameter values (e.g., a measurement error correlation of .10) leads to different values of practical fit indices when the incidental parameter values change. For example, it is well known that fit indices tend to indicate worse fit given the same amount of misspecified measurement error correlations when the measurement reliability on the indicators is high (Browne et al., 2002; Beauducel & Wittmann, 2005; Hancock & Mueller, 2011; Heene et al., 2011; Mahler, 2011; Savalei, 2012; Sharma et al., 2005). As a result, the one-size-fit-all cutoffs will reject the same misspecification in a CFA model (e.g., defined by a standardized cross loading of .2) with more reliable measures more often than in a CFA model with less reliable measures. This problem applies to not only misspecification in a measurement model but also misspecification in a structural model (Savalei, 2012). Saris et al. (2009) also showed that practical fit indices depend on the values of regression coefficients in multivariate regression model when the same amount of residual correlation is omitted.

1.2.2.4 Sample Size and Missing Data

The expected values of practical fit indices are not directly related to sample size although the shape of the sampling distribution of a fit index (e.g., variability) can vary across sample sizes. This is in fact a desirable property for effect size measures (Kelley & Preacher, 2012) because researchers can expect the same fit indices regardless of sample size. Previous studies (Beauducel & Wittmann, 2005; Davey, 2005; Hu & Bentler, 1999; La Du & Tanaka, 1989; Meade et al., 2008; Sharma et al., 2005) found that, using the same model with the same set of parameter values, sample size would not affect popular fit indices including RMSEA, CFI, and TLI (except for sample size less than 200). However, sample size will affect some of the fit indices that are not currently popular, such as GFI, AGFI (Adjusted Goodness-of-Fit Index), and RMR (Root Mean Squared Residual; Anderson & Gerbing, 1988; La Du & Tanaka, 1989). The expected values of the popular fit indices, however, are related to proportion of missing data, as well as the missing data patterns (Davey, 2005). I argue that a good effect size measure should not be sensitive to missing data. Future research is needed to develop practical fit indices that are not related to the presence of missing data.

1.2.2.5 Estimation Methods

Most fit indices are usually calculated from the chi-square values, which is the weighted combination of residuals. The chi-square values are influenced by the estimation method (Fan et al., 1999). Maximum likelihood (ML; and its modification), generalized least square (GLS), and asymptotic distribution free (ADF) methods are popular estimators for continuous data. Fit indices obtained from ML are different from GLS (Fan et al., 1999) and ADF (Schermelleh-Engel et al., 2003). Hu & Bentler (1999) limited their simulation to only ML so the derived cutoffs from Hu & Bentler (1999) cannot be applied to other estimators. In addition, the cutoffs cannot be used in data sets with different types of indicators (such as categorical data) that use different methods of estimation (Nye & Drasgow, 2011).

1.2.2.6 Sensitive to Data Distribution

ML, which is the most popular estimator, has an assumption that data must be multivariate normally distributed (Yuan & Bentler, 2004). If data are not normally distributed, the estimate of chi-square test statistic is inflated. Thus, Type I error is inflated and fit indices indicated worse fit. The transformation for reducing the impact of nonnormality on the central chi-square statistics is available (Bentler & Satorra, 2010; Satorra & Bentler, 2001). The scaled chi-square test statistic could control Type I error approximately equal to an alpha level.

When a hypothesized model is not a population model, the population fit indices are not perfect. Then, when data are normally distributed, the chi-square value is noncentral chi-square distributed with the noncentrality parameter defined in Equation 1.2. Note that when the noncentrality parameter is 0, the noncentral chi-square distribution will be a central chi-square distribution. In power analysis, the sampling distribution of the misspecified model is needed. When data are not normally distributed, however, the transformation for noncentral chi-square distribution is not available so the Type II error (1 - power) may be not valid (Yuan, 2005). In addition, the discrepancy value or noncentrality parameter estimate (see Equations 1.1 and 1.2) will be biased under nonnormal distribution. Consequently, the sample estimates of practical fit indices that are based on the discrepancy value are biased as well (Brosseau-Liard et al., 2012).

In conclusion, the primary benefit of fit indices is that they are effect size measures for model misfit. Fit indices, however, are sensitive to many factors, including model characteristics. Thus researchers should be aware of the influence of model characteristics in interpreting a fit index. For example, model size (or the number of indicators) should be considered when interpreting RMSEA. In the next section, I will describe additional problems when fit indices cutoffs are used.

1.3 The Problems of the Current Uses of Cutoffs on Fit Indices

Fit indices are not only used to quantify the degree of misfit of a hypothesized model. Cutoffs of fit indices are also developed to help researchers decide whether a hypothesized model sufficiently

fits the data. Currently, a one-size-fit-all cutoff has been frequently used for this decision regardless of model characteristics. I will illustrate three main problems of the current uses of fit indices.

1.3.1 The Problem of One-Size-Fit-All Cutoffs

As shown in the previous section, practical fit indices depend on multiple factors. Thus it is impossible to find an one-size-fit-all cutoff for any given fit index that is applicable to any models. For example, the cutoff provided by Hu & Bentler (1999) is based on a three-factor CFA model with 15 indicators. The standardized factor loadings in the CFA model ranged from .7 to .8. Their suggested cutoffs are limited to this specific model and this set of parameter values. In fact, Hu & Bentler (1999) provided the caution of using their suggested cutoffs in situations different from their testing models.

Although sample size, data distribution, or estimators are usually used as design factors in the simulation studies that derived the fit indices cutoffs, the same cutoffs do not provide the same Type I and II errors across sample size, data distribution, or estimators (Marsh et al., 2004). If researchers would like to control Type I or II errors to be equal across sample sizes, data distribution, or estimators, it is impossible to find the one-size-fit-all cutoff. Thus, model evaluation should not be based on the one-size-fit-all cutoff. Rather, the cutoffs should be adjusted based on the influencing factors described above.

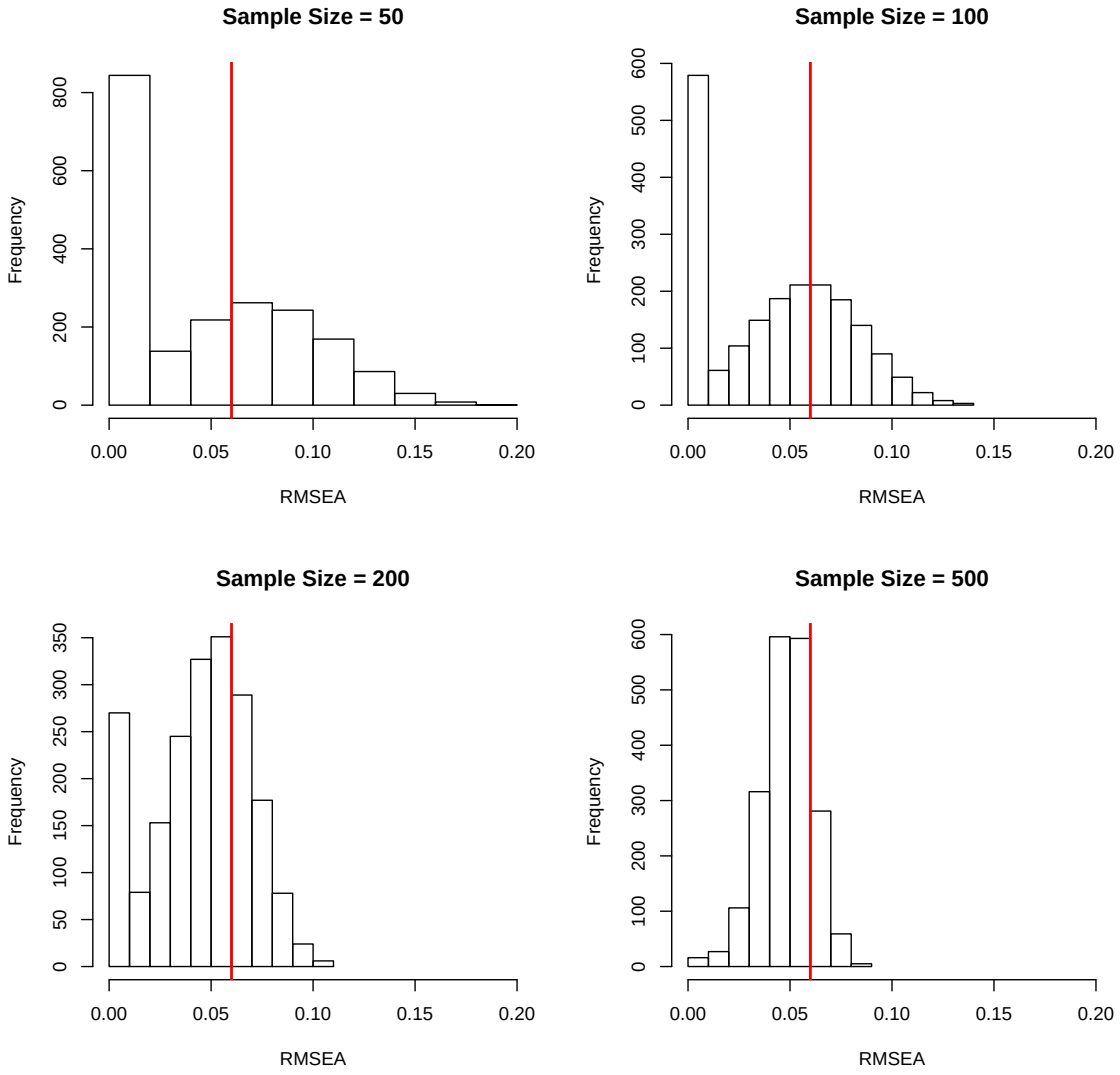
1.3.2 The Problems of the Derivation of the Cutoffs

Many simulation studies tried to derive the practical fit indices cutoffs or examine the performance of the derived cutoffs. These simulation studies used different principles to derive cutoffs. They also specified misspecified models by using different values of population fit indices. In addition, those simulation studies derived cutoffs by not considering the fact that the hypothesized models should be retained if they are only trivially misspecified.

1.3.2.1 Different Criteria Used for Deriving Cutoffs

Although the cutoffs facilitate decisions regarding whether to retain or reject hypothesized models, researchers do not know the Type I error rates resulted from the cutoffs. In fact, the same cutoffs can lead to different Type I error rates if model type or sample size changes. Marsh et al. (2004) suggested that Type I error should be controlled to a given alpha level and the statistical power (1-Type II error rate) to detect model misspecifications should increase as sample size increases like regular statistical testing. However, the fit indices cutoffs from Hu & Bentler (1999) were derived by minimizing both Type I and II errors. The derived cutoffs will not be the optimal cutoffs for all sample size conditions. Marsh et al. (2004) showed that the cutoffs suggested by Hu & Bentler (1999) increased Type II error rate (i.e., decreased power) as sample size increased. This result occurs when the population misfit (assuming that the misfit is nontrivial) is less than the suggested cutoffs. For example, the population RMSEA is .05 and the suggested cutoff of RMSEA is .06. As shown in Figure 1.1, the sampling error of RMSEA decreases when sample size increases. Then, the proportion of the areas below the cutoff (Type II error) is higher when sample size increases.

Figure 1.1: The problem of the increased Type II error when sample size increases. The population RMSEA underlying the sampling distributions is .05 assuming that this RMSEA value is based on a severely misspecified model. The RMSEA cutoff is .06. The rejection rate (statistical power) is lower when sample size increases. That is, the Type II error increases when sample size increases.



1.3.2.2 Different Population Practical Fit Indices for Misspecified Models across Simulation Studies

The past simulation studies used different types of models. The different types of models are needed to check whether suggested cutoffs are applicable to multiple types of models. The different types of models, however, resulted in different population fit indices. As a result, different

Table 1.1: Population Practical Fit Indices of Misspecified Models from Previous Simulation Studies that Tested the Sensitivity of Fit Indices or Fit Indices Cutoffs for Detecting Model Misspecification.

Studies	Count of Designs	Average of df	Average of ncp	Average of power	Average of RMSEA	Average of SRMR	Average of CFI	Average of TLI
Beauducel & Wittmann (2005)	16	438	25.787	.258	.026	.028	.919	.909
Curran et al. (2003)	9	55	16.262	.372	.051	.036	.964	.953
Davey (2005)	8	25	20.263	.462	.076	.083	.909	.867
Fan & Sivo (2007)	10	46	33.200	.671	.091	.086	.953	.928
Fan et al. (1999)	2	47	61.990	.884	.110	.036	.955	.937
Hancock & Mueller (2011)	12	129	51.388	.611	.057	.087	.930	.917
Heene et al. (2011)	12	515	81.798	.577	.049	.114	.845	.825
Heene et al. (2012)	8	251	164.511	.896	.083	.003	.992	.991
Hu & Bentler (1999)	4	80	38.473	.691	.065	.100	.945	.933
Jackson (2007)	36	165	21.971	.229	.026	.022	.981	.978
La Du & Tanaka (1989)	1	29	7.010	.230	.050	.025	.980	.967
Nye & Drasgow (2011)	2	90	19.710	.403	.046	.083	.946	.937
Schermelleh-Engel et al. (2003)	2	16	13.687	.550	.089	.087	.963	.935
Taylor (2008)	24	51	32.023	.447	.065	.043	.935	.916
Wu (2008)	51	9	6.415	.351	.077	.035	.987	.986

Note. Fit indices are based on the sample size of 100. df = degree of freedom, ncp = noncentrality parameter, RMSEA = root mean square error of approximation, SRMR = standardized mean square residuals, CFI = comparative fit index, TLI = Tucker-Lewis Index.

simulation studies suggested different cutoffs. Table 1.1 shows the average degree of misspecification in misspecified models used in 15 simulation studies. The Appendix A shows the degree of misspecification from all designs of each simulation study. I selected the simulation studies that defined misspecified models unless the studies explicitly said that the misspecification is trivial. To simplify this presentation, the simulation studies that have the same or similar designs as previous studies are combined in the table. For example, Fan & Sivo (2005) and Hu & Bentler (1998) are similar to Hu & Bentler (1999), and Chen et al. (2008) is similar to Curran et al. (2003). I realize that the list of simulation studies is not exhaustive but it is enough to show the heterogeneity of population fit indices. The degree of misspecification is defined in many ways: noncentrality parameter (Saris et al., 2009), statistical power to reject a model by chi-square test of absolute fit at sample size of 100 (Fan & Sivo, 2007), population RMSEA (Equation 1.3), SRMR (Equation 1.4), CFI (Equation 1.5), and TLI (Equation 1.6). Table 1.2 provides the descriptive statistics of the degrees of misspecifications across the 197 designs from the 15 studies.

As can be seen in Table 1.2, the population fit indices values substantially varied across studies. Some are lower and others are higher than the suggested cutoffs. For example, population RMSEA

Table 1.2: The Average Population Practical Fit Indices of 197 Misspecified Models from 15 Simulation Studies and the Average Values across All of the Studies.

Statistics	df	ncp	Power	RMSEA	SRMR	CFI	TLI
Across Designs							
Average	132.65	31.50	.421	.058	.047	.955	.944
Standard deviation	202.34	57.11	.343	.041	.050	.056	.068
Minimum	1.00	.00	.050	.000	.000	.635	.574
Maximum	945.00	462.17	1.000	.189	.312	1.000	1.000
Across Studies							
Average	129.54	39.63	.509	.064	.058	.947	.932
Standard deviation	155.30	40.33	.215	.024	.034	.037	.044
Minimum	8.71	6.42	.229	.026	.003	.845	.825
Maximum	514.50	164.51	.896	.110	.114	.992	.991

Note. Fit indices are based on the sample size of 100. df = degree of freedom, ncp = noncentrality parameter, RMSEA = root mean square error of approximation, SRMR = standardized mean square residuals, CFI = comparative fit index, TLI = Tucker-Lewis Index.

ranges from .025 to .110 at the study level. Sixty-three percent of the studies have RMSEA less than the suggested cutoff of .06 (Hu & Bentler, 1999). As a result, most designs in these studies will have decreased power to detect model misspecifications using the cutoff when sample size increases, as shown in Figure 1.1. If the rule of minimizing both Type I and II errors is used for all studies, this heterogeneity of population fit indices will surely lead to different suggested cutoffs for any fit indices.

1.3.2.3 The Derivation Excluding Trivial Misspecification

One reason why practical fit indices are preferred to chi-square test statistic is that researchers wish to retain models with only trivial misspecifications. That is, if misspecification in a model is negligible, researchers would like to claim that their hypothesized model approximately fits data well. The negligible misspecified parameters are sometimes referred to as parsimony error. Almost all past simulations, however, calculated Type I error based on models with perfect fit. That is, they created population models without imposing any parsimony error (except Cheung & Rensvold, 2002) in spite of the fact that researchers allow trivial misspecification in practice. As

a result, the cutoffs are derived by minimizing the rejection rate of a perfect-fitting model. Rather, the cutoffs should account both parsimony and sampling errors (Cheung & Rensvold, 2001). The cutoffs should be based on Type I error that is defined as the proportion of falsely rejecting model with approximate fit.

1.3.3 The Problem of the Divergent Results from Different Indices

Marsh et al. (1988) showed that more than 30 fit indices were available at the time of study. Multiple fit indices provide information on different aspects of model fit. For example, SRMR is sensitive when nonzero factor correlations are misspecified as 0 but RMSEA is sensitive to the misspecification on cross loadings (Hu & Bentler, 1998). Thus, researchers could get inconsistent information from multiple fit indices. In the worst scenario, researchers may intentionally use the fit indices indicating good fit and ignore the fit indices providing bad fit to support their hypotheses. Thus, a method to combine the information from different indices is needed. As the first attempt, Hu and Bentler (1998; 1999) proposed a two-index strategy which combines SRMR with a chi-square based fit index (e.g., RMSEA or CFI) for model fit evaluation. However, this strategy is criticized by Fan & Sivo (2005) which showed that it was only valid in limited circumstances. Thus, to date there is no good way to combine the different fit indices. This article provides a unified approach that combines information from different fit indices together.

In conclusion, it is unrealistic to find a fixed one-size-fit-all cutoff for a fit index that would apply to any models or sample sizes is unrealistic (Chen, 2007; Chen et al., 2008) given that there are multiple influencing factors on practical fit indices. Those influencing factors are the main reasons why there is no consensus on the suggested cutoffs. Therefore, rather than spending the effort on finding a one-size-fit-all cutoff, the methods of model evaluation accounting for those influencing factors are needed.

Chapter 2

Alternative Approaches

Many alternative approaches for model evaluation have been developed to solve the problems of the use of one-size-fit-all cutoffs to evaluate model fit. A common advantage of the alternative approaches is that they systematically account for both model parsimony and sampling errors, although model parsimony is defined in different ways.

2.1 Test of Close Fit and Not Close Fit

The chi-square test statistic is used to test the null hypothesis that the hypothesized model perfectly represents relationship in the data. This null hypothesis is not consistent with the goal of using fit indices to identify models with approximate fit to a population model. Browne & Cudeck (1992), MacCallum et al. (1996), and MacCallum et al. (2006) argued that users may define a new null hypothesis that focuses on the approximate fit of a hypothesized model. The approximate fit can be defined by a small population RMSEA value (e.g., .05) which indicates a trivial difference between the hypothesized and population models. The population RMSEA value can be used to estimate a noncentrality parameter which defines a noncentral chi-square distribution. The sampling distribution of the sample RMSEA can be then derived based on the noncentral chi-square distribution. With the sampling distribution derived, researchers have two options in testing model fit: test of close fit and test of not close fit.

The test of close fit is to test whether a hypothesized model provides a worse fit to the population underlying the data than the specified approximate fit. For example, the null hypothesis to be tested may be that the population RMSEA is less than or equal to .05 ($H_0 : \varepsilon \leq .05$). When the null hypothesis is rejected, researchers would conclude that population RMSEA is larger than .05, indicating that the hypothesized model does not provide an approximate fit to the population model. Note that failure to reject the null hypothesis does not necessarily indicate that the hypothesized model is trivially misspecified. It could be the case that users do not have enough statistical power to reject the null hypothesis of close fit due to small sample size. Alternatively, test of not close fit is to test whether the hypothesized model provides a better fit to the population underlying the data than the approximate fit. For example, researchers may test the null hypothesis that the population RMSEA is larger than or equal to .05 ($H_0 : \varepsilon \geq .05$). Rejecting the null hypothesis would indicate that the hypothesized model provides an approximate fit to the population model.

Conceptually, the test of not close fit is equivalent to equivalence testing (Seaman & Serlin, 1998; Steiger, 2004). Equivalence testing is used to test whether two target statistics (e.g., group means) are approximately equivalent (i.e., trivially different). Equivalence testing is popular in biostatistics research in which researchers wish to test whether two treatments are equally efficient. For example, in comparing sample means, researchers will test two null hypotheses: $H_0 : \mu_1 - \mu_2 \geq \delta$ and $H_0 : \mu_1 - \mu_2 \leq -\delta$, where δ is the maximum level of trivial difference (Schuirmann, 1987). If both null hypotheses are rejected, researchers can claim that two population means are trivially different (or equivalence). The same idea is used by SEM users to test whether the population model underlying the data and the hypothesized model are approximately equivalent.

Test of close fit or not close fit might also be done simultaneously using the confidence interval of RMSEA (Westlake, 1976; Kirkwood & Westlake, 1981). If both lower and upper bounds of the confidence interval are lower than the maximal trivial misfit value of RMSEA (e.g., .05), the null hypothesis in the test of not close fit is rejected. In other words, the amount of misfit is trivial. If both lower and upper bounds of the confidence interval are higher than the RMSEA, the null hypothesis in the test of close fit is rejected. That is, the amount of misfit is not trivial.

If the confidence interval brackets the RMSEA, the result is inconclusive. Note that the tests of close fit and not close fit are both one-tailed tests. Let α represent the alpha level of a one-tailed test. $100(1-2\alpha)\%$ confidence interval should be used (Steiger, 2004). Note that this procedure is similar to the significant testing proposed by Jones & Tukey (2000) resulting in three options: model misfit greater than the maximal trivial misfit, model misfit less than the maximal trivial misfit, or inconclusive.

In this approach, model parsimony is accounted for by population RMSEA value of maximal trivial misfit and sampling error is accounted for by the sampling distribution of RMSEA. However, there are several limitations of this approach. First, the chi-square test statistic is assumed to follow a noncentral chi-square distribution. This assumption is only plausible when data are multivariate normal distributed and the amount of misfit is not large (Yuan & Bentler, 2004; Yuan, 2005). Second, this approach is limited to fit indices whose sampling distributions can be analytically derived, such as RMSEA. Furthermore, the definition of parsimony error is based on population value of fit indices. As described before, the population fit indices are sensitive to model characteristics, such as model size, type of models, and incidental parameters.

2.2 Modification Indices and Power Approach

Saris et al. (2009) proposed to use modification indices (MI) and the power of MI to test potential nontrivial misspecifications for a hypothesized model. This method involves several steps. First, researchers need to specify a maximally acceptable degree of misspecification for all fixed parameters of a hypothesized model. For example, Saris et al. (2009) proposed that omitted cross loadings (which are constrained to be 0 in a hypothesized model) should not be greater than .4 in a standardized metric and omitted regression paths should not be greater than .1 in a standardized metric to be deemed as trivial. The power of detecting the maximally acceptable misspecification is then computed for each parameter.

Second, for each fixed parameter, significance of MI (no vs. yes) is obtained, which is ex-

amined along with the power of detecting maximally acceptable misspecifications (low vs. high), resulting in four combinations. (a) If the power is low and the MI is significant, the misspecification is not trivial. (b) If the power is high and the MI is not significant, the misspecification is trivial. (c) If the power is high and the MI is significant, then the expected parameter change (EPC) when the parameter is freed should be examined. If EPC exceeds the range of maximally acceptable degree, the misspecification is not trivial. Otherwise, the misspecification is trivial. (d) If the power is low and the MI is not significant, the decision is inconclusive because of low power in detecting nontrivial misspecification.

The advantage of the approach is that it provides information besides the significance of MI. The fixed parameter with a significant MI is usually viewed as severely misspecified even though the EPC does not exceed the level of maximal trivial misfit. This approach requires researchers to define maximally acceptable misspecification (i.e., parsimony error) in fixed parameters of a hypothesized model and to investigate whether EPCs exceed the defined parsimony errors. The definition is limited to fixed parameters, however. Some types of misspecification are not allowed by this approach, such as an omitted factor explaining additional relationship among indicators. As another limitation, combining the significance testing of MI and the power to detect maximal trivial misspecification does not accurately account for the sampling error of EPC. I will show two example scenarios that this framework provides inconsistent results.

In both scenarios, the misspecified parameter is a cross loading for which -.4 and .4 are maximal trivial misspecified value. Suppose the EPC of a cross loading is .35. In Scenario 1, let the standard error of the EPC of the cross loading be .153, which will lead to a significant MI. The power to detect maximal trivial misspecification in this scenario is 74%, which is not high enough regarding the criterion of .8. Because the observed EPC is .35 with a significant MI and low power, the parameter is misspecified according to Saris et al. (2009) (see combination a).

In Scenario 2, let the standard error of the EPC be 0.10, which will lead to a significant MI. The power in detecting maximal trivial misspecification is 98% indicating high power. According to Saris et al. (2009) (see combination c), the parameter is trivially misspecified because the observed

EPC is less than the maximal trivial misspecified value of .4. Note that the estimated EPC in both scenarios are equal and less than the maximal trivial misspecified values. According to Saris et al's guidelines, one scenario indicates trivial misspecification but another scenario indicates severe misspecification.

Rather than using the combination of the significance of MI and the power in detecting maximal trivial misspecified values, the confidence interval for the EPC provides the magnitude of EPC accounting for sampling error. Researchers can examine whether the confidence interval brackets the maximal trivial misspecified values. The Wald confidence interval of EPC (θ) can be calculated by Equation 2.1:

$$CI_{1-2\alpha}(\theta) = \{\theta | EPC - z_{\alpha} SE_{\theta} < \theta < EPC - z_{1-\alpha} SE_{\theta}\}, \tag{2.1}$$

where z_{α} is the α quantile of the standard normal distribution and

$$SE_{\theta} = EPC/\sqrt{MI}. \tag{2.2}$$

The equivalent testing framework is also applicable for EPC. A 90% confidence interval is used for the alpha level of .05. For Scenario 1, the 90% confidence interval is 0.10 to 0.60. The EPC is not zero but it ranges from a trivial (0.05) to a large value (0.65). Thus the decision here should be inconclusive. For Scenario 2, the 90% confidence interval ranges from a trivial (0.19) to a large value (0.51) so the decision should be inconclusive as well. Thus, both examples show that the use of CI for EPC is more informative than the combination of significance test of MI and the power and lead to more consistent results.

2.3 Bayesian Analysis

The Bayesian estimation approach allows users to simultaneously estimate target parameters and fixed parameters such as cross loadings. I will use "nontarget parameters" instead of "fixed param-

eters" in the Bayesian analysis because these "fixed parameters" (usually fixed as 0) in maximum likelihood estimation (MLE) are estimated in the Bayesian analysis. Nontarget parameters are usually fixed as 0 in MLE for the model identification purpose although they may actually have small values in the population. The Bayesian analysis will allow these nontarget parameters to be estimated by imposing informative prior distributions (Muthén & Asparouhov, 2012). For example, cross loadings may be given normal priors with the means of 0 and the standard deviations of 0.1. This normal priors will result in 95% CIs for the cross loadings between -.2 and .2. Note that the prior distributions of target parameters (e.g., hypothesized factor loadings) will also need to be specified. The prior distributions can be noninformative (e.g., normal distribution with very wide standard deviation) or informative.

After specifying the priors, the Bayesian approach will combine the information from the specified priors and the likelihood of parameters given the observed data, which will result in posterior distributions for both target and nontarget parameters. Posterior distributions show the plausible ragnes of prameters given the observed data. The nontarget parameters will be highly influenced by informative priors so the posterior distributions will be close to the range of values specified in prior distributions.

To evaluate model fit, analysts need to investigate global and local fits simultaneously. Global fit can be evaluated by Posterior Predictive P-value (PPP; Gelman et al., 1996; Levy, 2011). Multiple sets of parameter values are drawn from the posterior distribution. PPP is computed by comparing the likelihood of observed data and the likelihood of data simulated from each drawn set of the posterior parameter values. PPP is the proportion of the likelihood of observed data that is less than the likelihood of simulated data across all drawn sets of the posterior parameter values. If a model fits well, PPP is close to .5. A low PPP value implies model misfit. PPP does not have the same interpretation as the traditional *p*-value in inferential statistics and PPP is viewed as a practical fit index. Muthén & Asparouhov (2012) indicated that .05 cutoff similar to traditional *p*-value appeared reasonable.

Muthén & Asparouhov (2012) also suggested researchers investigate the posterior distributions

of nontarget parameters to evaluate local fit. If the span of a posterior distribution (referred to as credible interval in the Bayesian framework, e.g., middle 95%) does not cover 0, the nontarget parameter is misspecified. In fact, the maximal misspecified parameter values proposed by Saris et al. (2009) and the equivalence testing can be applied here. Researchers can also investigate whether the upper and lower bounds of the credible intervalis are over the maximum acceptable values of parameters. If so, nontarget parameters are severely misspecified.

The Bayesian approach is able to not only take into account trivial model misspecification at the parameter level, but also estimate the trivial model misspecification. However, not all possible trivial model misspecifications can be accounted for at once in a hypothesized model; otherwise, the Bayesian estimator is not able to converge. For example, either cross loadings or residual covariances are used as trivial model misspecification (except putting informative priors to target parameters in hypothesized model).

Moreover, the results of local fits depend on sample sizes and specified priors (Jorgensen et al., tion). A misspecified nontarget parameter will be more likely to cover 0 when a sample size is smaller and specified priors are narrower. Smaller sample sizes will provide wider credible interval so the coverage rate of 0 is higher. Narrower priors make the estimated nontarget parameters closer to 0 so the coverage rate of 0 is higher.

The expected values of PPP depend on sample size and specified priors controlling the magnitude of misspecified parameters (Jorgensen et al., tion). In a severely misspecified model with the narrow priors (normal priors with *SD* of 0.005), PPP is slightly smaller than .50 with small sample sizes and approaches 0 with larger sample sizes. In severely misspecified model with the wide priors (normal priors with *SD* of 0.1), PPP is slightly smaller than .50 with smaller sample sizes and approaches .50 with larger sample sizes. With the medium priors (e.g., normal priors with *SD* of 0.05), PPP approaches any values between 0 and .50 with larger sample sizes. Because PPP depends on sample size and priors, the one-size-fit-all cutoff of PPP is problematic. Jorgensen and colleagues showed that Type I and II errors based on the suggested cutoffs (Muthén & Asparouhov, 2012) were different across sample sizes.

Thus, model fit evaluation with the Bayesian approach is highly sensitive to data and specified priors. This sensitivity to priors can be viewed as both a strength (accounting for analysts beliefs) and a weakness (no consensus on specifying priors) of Bayesian estimation.

2.4 Simulation Approach

Millsap (2007; 2010; 2013) and Millsap & Lee (2008) proposed a simulation approach to establish a sampling distribution of a fit index if trivial misspecification presents. This method is an improvement from parametric bootstrap (Efron & Tibshirani, 1993) by adding parsimony error into the sampling distribution. In this approach, the parameter estimates from fitting a hypothesized model to observed data are used to simulate a large number of data sets. The simulated data sets are then fitted by the hypothesized model to obtain the empirical sampling distributions of fit indices. Millsap suggested adding user-defined parsimony error at the parameter level so that the hypothesized model is not equal to but still a good approximation of the population model. For example, the values of two fixed cross loadings are changed from 0 to .3 in standardized scale. Then, the modified parameter estimates are used in data generation. By generating data from the population model with parsimony error, the sampling distributions would reflect both parsimony and sampling errors. Then, a plug-in p value is computed as the proportion of fit indices from the empirical sampling distribution that indicates worse fit than the fit index value obtained from the observed data (Robins et al., 2000). The plug-in p value is compared with a priori alpha level (e.g., .05). If the plug-in p value is greater than the alpha level, the hypothesized model is a good approximation of the population underlying the data. Otherwise, one can then conclude that the hypothesized model cannot approximate the population behind the original data with a priori definition of trivial misspecification.

The rationale of this method is similar to the chi-square test and the test of close fit mentioned above. But the difference is that the sampling distribution will be established empirically instead of analytically. Furthermore, this method can account for nonnormal data distribution by the data

generation using Bollen-Stine bootstrap (Bollen & Stine, 1992; Millsap, 2013).

The main problem of the simulation approach is that it assumes that failing to reject the null hypothesis of approximate fit indicates approximate fit. This is the same problem of accepting a null hypothesis when researchers fail to reject the null hypothesis. Researchers may have low power to detect severe misspecification. Rather, the test of not close fit (i.e., equivalence testing) should be used for the evidence of approximate fit. Furthermore, the simulation approach allows users to specify only one set (or a limited number of sets) of parsimony error. There are infinite ways to specify parsimony error, such as having cross loadings in a different set of indicators. All possible user-defined sets of parsimony error should be covered in the model evaluation as much as possible. Later, the unified approach will provide a method to account for multiple sets of parsimony error.

The four alternative methods presented above account for parsimony error which require users to define model parsimony, either as misspecified parameter values or population practical fit indices. No method is perfect. They all have their own advantages and limitations. However, they can complement each other so that more accurate inference for model fit evaluation can be made. In the next chapter, I will propose a unified approach which tends to combine all alternative methods (except Bayesian framework) into a single framework to utilize all of their advantages.

Chapter 3

The Unified Approach

In this section, the test of close fit / not close fit, the modification indices approach, and the simulation approach are integrated into a single framework referred to as the unified approach. This unified approach still utilizes fit indices because they are useful effect size measures. However, the unified approach does not use one-size-fit-all cutoffs of the fit indices, but it derives fit indices cutoffs corresponding to a specified Type I error and specific model and data characteristics.

As mentioned in Chapter 1, there are three primary problems of the use of fit indices cutoffs. The unified approach is designed by integrating three alternative approaches mentioned above so that it can address most of the problems. The unified approach takes model characteristics and sample size into account so the results of the unified approach should not depend on model characteristics or sample size. In addition, the unified approach provides rules to combine information from different fit indices together so it prevents the problem that different fit indices provide different conclusions. However, it does not completely solve the subjectivity problem in the derivation of the fit indices cutoffs. Researchers need to define trivially and severely misspecified models for which the definition could be subjective. However, the benefit of the unified approach is that it encourages researchers to explicitly clarify trivially and severely misspecified models in the scale that most researchers can understand, such as the magnitudes of factor loadings or error correlations, instead of the scale of a fit index. Researchers rarely know the meaning behind a cutoff value

of a fit index. For example, researchers do not know the magnitude of misspecified cross loadings when RMSEA is .05. This problem will be discussed further in Sections 3.3 and 6.3.3.

The unified approach consists of two parts: global fit evaluation and local fit evaluation. The global fit evaluation focuses on finding the range of values for each of fit index when a model is trivially or severely misspecified. The local fit evaluation focuses on all fixed parameters in the model by using confidence intervals of EPCs. The conclusion from both global and local fit evaluation can be trivial misspecification, severe misspecification, or inconclusive. The results from the two parts are combined together to provide a single result, which is trivial misspecification, severe misspecification, or inconclusive.

3.1 Global Fit Evaluation

For global fit evaluation, the process of the unified approach is similar to the test of close fit / not close fit or the simulation approach. The global fit evaluation method is divided into multiple steps.

3.1.1 Step 1: Specify a Hypothesized Model and its Parameter Values

Researchers need to specify the parameter values in a hypothesized model which are used to simulate data and quantify trivial misspecifications. The parameter values may come from theory (e.g., high factor loadings from highly reliable measures), previous studies, or parameter estimates from observed data (similar to parametric bootstrap). The parameters in the hypothesized model are referred to as target parameters. The parameter values are referred to as target population values.

3.1.2 Step 2: Specify Parsimony Errors

Parsimony error can be defined at the parameter level (e.g., omitting an error correlation of .1) or at the global fit level (e.g., population RMSEA of .05). Parsimony errors are imposed to the hypothesized model from Step 1 so that data are generated from a model with approximate fit to

the hypothesized model. The hypothesized model with target parameters and parsimony errors is referred to as an alternative model (Millsap, 2013).

If parsimony error is specified at the parameter level, users may (a) pick a set of fixed parameters (e.g., two cross loadings), (b) change the fixed parameters to a value that represents the maximal trivial misspecification (e.g., .2 for standardized cross loadings). This method is used in the simulation approach (Millsap, 2013). I will refer this method to as a fixed parameter method. Suppose the hypothesized model is a CFA model without correlated unique factors. Two covariances among unique factors are included so that the values are equivalent to the measurement error correlations of .1. The limitation of the fixed parameter method is that users can only specify one or a few types of maximal trivial misspecification even though there are countless types of trivial misspecifications for a hypothesized model. The results of the model evaluation would highly depend on the selection of the modified fixed parameters.

Users may specify parsimony error at the population-fit level, which is referred to as the population misfit method. The population misfit method does not bind with any specific parameters or any types of misspecification. Infinite sets of misspecification at the parameter level can provide the same value of population fit index. Users simply specify a value of population fit index to account for all possible misspecifications that lead to the same amount of misfit. Population fit indices, however, depend on model characteristics. The maximal trivial misspecification defined by population fit indices need to be redefined once the hypothesized model changes. Unfortunately, researchers rarely know what population fit index values would be appropriate to their hypothesized models. Furthermore, different sources of misspecification resulting in the same amount of misfit may be resulted from different degrees of misfits at the parameter level. For instance, Fan & Sivo (2005) showed that constraining a factor loading of .455 as 0 (standardized loading of .376) would result in an equal population misfit as fixing a factor correlation of .50 as 0. However, a standardized factor loading of .376 is considered trivial in Saris et al. (2009) but a factor correlation of .50 is considered as medium effect in Cohen (1988).

Therefore, I recommend to specify parsimony error at the parameter level and account for mul-

tiple forms of maximal trivial misspecifications. I propose a new method to accommodate multiple sets of trivial parameters instead of one set of trivial parameters. This method is referred to as a repeated sampling method which defines trivial parameters as a range of misspecified fixed parameters. For example, the magnitude of standardized cross loadings less than .2 is deemed trivial. The repeated sampling method tends to find a combination of trivial parameters that maximizes misfit. For this example, cross loadings can be randomly drawn from uniform distributions of -.2 to .2 resulting in a combination of cross loading values. The random process is repeated for many times (e.g., 1,000). Population fit indices are calculated for each combination. The combination providing the maximum misfit (referred to as maximal trivial misspecification) is used for data generation. The obtained global misfit is a maximum value of misfit that could be considered as trivial misfit. The limitation of this approach is that the combination with the real maximal misfit is unlikely to be picked. The real maximal misfit combination may be not drawn. The more combinations being compared, the closer the level of misfits to the real maximal misfit.

Different population fit indices, however, are sensitive to different types of trivial parameters. For example, RMSEA is sensitive to an omitted cross loading but SRMR is sensitive to an omitted factor correlation (Hu & Bentler, 1999). Therefore, different fit indices may lead to different maximal trivial misspecification. For example, if RMSEA, CFI, TLI, and SRMR (calculated by Equations 1.1-1.6) are used to define population misfits, there will be up to 4 combinations of maximal trivial misspecifications to be used in data generation in the next step.

In conclusion, target parameter values with parsimony error are specified during this step. The parsimony error can be defined by (a) the population misfit method, (b) the fixed parameter method, or (c) the repeated sampling method. Because the repeated sampling method solves the problems of the population misfit method and the fixed parameter method, I will focus only on the repeated sampling method in the following steps.

3.1.3 Step 3: Account for Sampling Error

As illustrated above, different sets of maximal trivial misspecification may be identified by using different fit indices. In this step, each set of maximal trivial misspecification is used to generate data and establish the sampling distribution of the corresponding fit index. For example, if a set of maximal trivial misspecification is identified using RMSEA, this set is then used to establish the sampling distribution of RMSEA. This sampling distribution accounts for both maximal parsimony error and sampling error. Researchers may generate data by multivariate normal distribution or the Bollen-Stine bootstrap (Millsap, 2013). The Bollen-Stine bootstrap is used to account for nonnormal data distribution. The Bollen-Stine bootstrap transform the observed data so that the transformed data represent the hypothesized model with specified parsimony error. The data are transformed by the following formula (Bollen & Stine, 1992; Yung & Schumacker, 1996):

$$\boldsymbol{Z} = \left[\boldsymbol{Y} - \mathbf{1}\hat{\boldsymbol{\mu}}'\right]\hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}_{MTM}^{1/2} + \mathbf{1}\boldsymbol{\mu}'_{MTM}, \tag{3.1}$$

where $\boldsymbol{Z}$ is the transformed data, $\boldsymbol{Y}$ is the observed data, $\mathbf{1}$ is a unit vector with a length equal to sample size, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are sample estimates of mean vector and covariance matrix, and $\boldsymbol{\mu}_{MTM}$ and $\boldsymbol{\Sigma}_{MTM}$ are model-implied mean vector and covariance matrix from a hypothesized model with a maximal trivial misspecification. The transformed data are then resampled with replacement. The hypothesized model is fitted to the resampled data to obtain the empirical sampling distribution of a fit index given a specified parsimony error.

3.1.4 Step 4: Examine whether hypothesized model are Severely Misspecified

Observed values for the fit indices are obtained from fitting the hypothesized model to observed data. The observed fit indices are compared with the corresponding sampling distributions established in Step 3. Similar to hypothesis testing, critical region or plugin p-value can be used to decide whether the observed fit indices are likely from the population behind the sampling distri-

butions. The critical region is the area spanning $100\alpha\%$ on the poor-fit side (e.g., higher values for RMSEA). If an observed fit index falls in the critical region, the hypothesized model is deemed severely misspecified. Otherwise, the hypothesized model could be trivially or severely misspecified. Steps 5-7 will show the reasons why the model can be severely misspecified when the observed fit indices are better than the critical values. Note that this decision is different from the simulation approach that the model is deemed trivially misspecified when researchers fail to reject the null hypothesis. The plugin p-value is the proportion of the generated data sets showing poorer fit than the observed fit index obtained from the observed data. If the plugin p-value is less than the alpha level, the model is rejected. Otherwise, the model could be trivially or severely misspecified. This procedure is similar to the test of close fit. The test is used to investigate whether the misfit of the hypothesized model to the observed data are worse than the misfit due to maximal trivial misspecifications. This step is repeated for all fit indices of interest.

Sometimes, the hypothesized model is rejected by some practical fit indices but not the others. If one fit index suggests that the model should be rejected, then the model should be rejected regardless of the results of the other fit indices. The reason is that each fit index capture different types of misfit. Then, a severely misspecified model can indicate model misfit in one fit index whereas the other fit indices do not indicate model misfit. If all fit indices do not suggest model rejection, then the hypothesized model could be trivially or severely misspecified.

3.1.5 Step 2a-4a: Shortcut to Examine whether Hypothesized Models are Trivial

Testing whether observed fit indices are significantly worse than maximal trivial misspecifications can be time consuming because it involves finding the sets of maximal trivial misspecifications (by the repeated sampling method) and fitting a hypothesized model to the generated data sets. These steps can be shorten by the following inequalities. Let δ be a population fit index that a lower value indicates better fit. Let $\delta_1, \delta_2, \ldots, \delta_k$ be the population fit indices calculated from the draws from the repeated sampling method where k is the number of draws. If δ_{MTM} is the maximal

trivial misspecification defined by δ, then $\delta_{MTM} \geq \delta_1, \delta_2, \ldots, \delta_k$. Let δ_C be the critical value from empirically deriving sampling distribution from δ_{MTM}. Assume that the sampling distribution of δ is not extremely negatively skewed. Then, $\delta_C \geq \delta_{MTM}$. If the sample estimate of δ, $\hat{\delta}$, is less than any of $\delta_1, \delta_2, \ldots, \delta_k$, then $\delta_C \geq \hat{\delta}$. Hence, when an observed fit index has a better fit than the fit index from any draws in the repeated sampling method, a model is failed to reject.

3.1.6 Step 5: Find Minimal Severe Misspecification

A good hypothesized model should have the amount of misfit that does not exceed both maximum trivial misspecification and minimum severe misspecification levels. In specifying misfit in one dimension, the thresholds for maximum trivial misspecification and minimum severe misspecification are identical. For example, suppose a difference of 0.2 in two-group latent factor means (says that the standard deviations of latent factors are 1) is deemed a trivial difference. The difference greater than 0.2 is then nontrivial. So 0.2 serves as a threshold for both maximum trivial misspecification and minimum severe misspecification. However, when there is more than one dimension, researchers need to aggregate the misfit from multiple dimensions into a single index (i.e., fit index measure). The thresholds would no longer be the same. The minimum severe misspecification may indicate less amount of misfit than the maximum trivial misspecification. For example, let the latent mean difference of 0.2 be the threshold for maximum trivial misspecification for each comparison. Researchers would like to investigate the difference in two factors. The trivially-misspecified model could be that both factors have the difference of 0.19 across two groups. The severely-misspecified model could be either of two factors has the difference of 0.21 across two groups. The latter situation usually has a less population misfit less that the former situation.

Therefore, instead of considering only maximum trivial misspecification, researchers should consider both maximum trivial misspecification and minimum severe misspecification if the dimension of misfits is greater than 1. Researchers can use the repeated sampling method to pick the point of maximum trivial misspecification described above. Researchers should be careful of using

the repeated sampling approach with minimum severe misspecification. Some sets of misspecification can be reparameterized into the estimated parameters and the amount of misfit can be zero. For example, if all target factor loadings are equal and all misspecified cross loadings are specified with the same value, the misfit could be 0 because this model can be reparameterized as the hypothesized model with different values of target loadings and factor correlations (i.e., factor indeterminacy, see Savalei, 2012). Therefore, I recommend users to randomly pick one dimension of misfit and specified at the threshold of maximal trivial misspecification. For example, the minimal severe misspecified model can be attained by randomly picking one cross loading from all cross loadings and put the thresholds, such as .4 or -.4. The set that provides the minimum value of misfit is then used as the minimal severe misspecification. Researchers may find the minimal severe misspecification by listing all possible combinations with one dimension of misfit and comparing the level of misfit. For example, one factor model with ten indicators has 45 ($10 \times 9/2 = 45$) possible omitted measurement error correlation. Researchers may set the omitted error correlations as .1 or -.1. Then, researchers can compare the level of misfits from all 90 combinations ($45 \times 2 = 90$) and pick the combination with the minimum level of misfit. Note that, if multiple fit indices are used to quantify different dimensions of population misfits, different fit indices may provide different sets of minimum severe misspecification.

3.1.7 Step 6: Account for Sampling Error in Severe Misspecification

Different sets of minimal severe misspecification are defined by different fit indices. In this step, each set of minimal severe misspecification is used to generate data and find the sampling distribution of the corresponding fit index. Data generation can be based on multivariate normal distribution or the Bollen-Stine bootstrap approach provided in Equation 3.1.

3.1.8 Step 7: Examine whether Observed Fit Indices are Trivially Misspecified

Observed fit indices are compared with the sampling distributions of each fit index given minimal severe misspecifications. The observed fit indices are tested whether they have significantly better fit than minimal severe misspecification (which will be better maximal trivial misspecificaiton). The critical region would be the area spanning $\alpha\%$ on the better-fit side (e.g., lower values for RMSEA). The plugin *p*-value is the proportion of the generated data sets having better fit that the fit index obtained from the observed data. If the test is significant, the model has a trivial (or no) misspecification. If the test is not significant, the model could be trivially or severely misspecified.

Again, the results across different fit indices may be inconsistent. To claim that the model is trivially misspecified, all fit indices should suggest trivial misspecification. The reason is that all fit indices provide different information of misfit. The trivial misspecification can be claimed if severely misspecified model are unlikely to be true in all aspects of misfit.

Based on the tests with maximal trivial misspecifications and minimal severe misspecification, there are three possible outcomes: severe misspecification, trivial (or no) misspecification, and inconclusive. If the result is inconclusive, more information can be acquired from local fit evaluations described later.

3.1.9 Step 5a-7a: Shortcut to Examine whether Observed Fit Indices are Trivial

Similar to testing with maximal trivial misspecification, testing with minimal severe misspecification can be shorten. Using the same logic as Step 2a-4a, when an observed fit index has a worse fit than the fit index from any draws in the repeated sampling method, a model cannot be claimed as trivially misspecified.

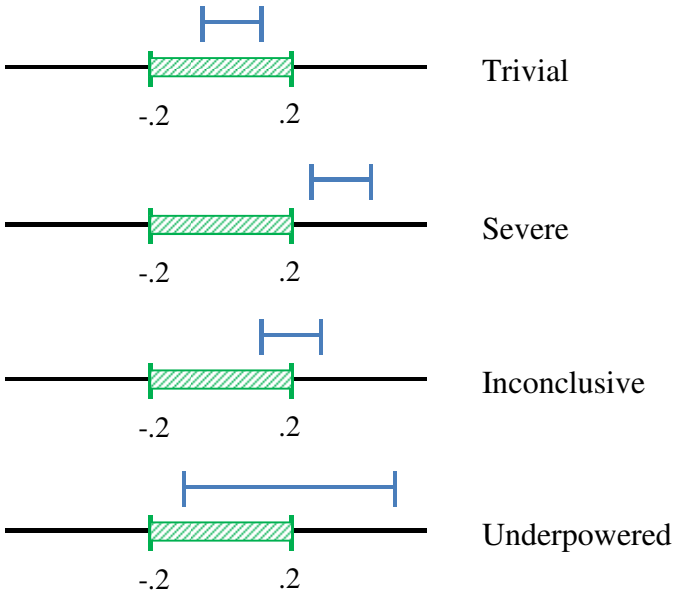
3.2 Local Fit Evaluation

Local fit evaluation is the investigation whether each constraint in a model is tenable. When a hypothesized model does not fit observed data well, researchers usually search for any potential local misfits. Also, when the global fit evaluation provides inconclusive results, researchers may evaluate local fit to gain additional information. The local misfits can be searched by using EPC. Currently, researchers check whether the EPCs (i.e., modification indices) are significantly different from 0. This approach does not acknowledge that a fixed parameter can be trivially misspecified. Saris et al. (2009) proposed how to account for trivial misspecification in using MI but their approach is problematic as described above. This section propose a way to evaluate EPC using confidence intervals.

The same paradigm for global fit is applicable for local fit. That is, researchers specify a level of maximal trivial misspecification. Because local misfit is based on only one dimension of misfit, the minimal severe misspecification will be the same point as the maximal trivial misspecification. Researchers may derive the sampling distributions of EPC to see whether the null hypotheses based on test of close fit or not close fit are rejected. The test of close fit and not close fit can be indirectly tested using confidence intervals. The $100(1-2\alpha)\%$ confidence interval (Equation 2.1 for EPC) should be used because it is a one-tailed test (e.g., use 90% confidence interval for $\alpha = .05$). Then, if both lower and upper bounds of the confidence interval are in the range of maximal trivial misspecification (e.g., a value between -.2 to .2 for a standardized cross loading), the model is trivially misspecified. If the confidence interval brackets the maximal trivial misspecification values, the decision is inconclusive. If the confidence interval does not overlap with the range of trivial misspecification, the fixed parameter is severely misspecified. If the width of confidence interval is wider than the range of trivial misspecification (e.g., the width of .4 for the trivial misspecification ranged from -.2 to .2), lower and upper bounds cannot be both inside the range of trivial misspecification so one cannot decide whether the misspecification is trivial. This is probably due to the fact that the test is underpowered. See Figure 3.1 for a graphical illustrations. In sum, one can obtain four outcomes for each fixed parameter: severely misspecified, trivially misspecified,

inconclusive, and underpowered. Note that users may define their trivial misspecification in standardized scales but the EPC is in the unstandardized scale. Analysts must transform either value to make them comparable.

Figure 3.1: The decisions for the confidence interval of an expected parameter change. The green bands represent the range of trivial misspecification. The blue lines represent the 90% confidence interval of an expected parameter change.



If the global fit evaluation is inconclusive, the information from the local fit evaluation may be used to infer global fit. EPC tests on multiple fixed parameters can provide different results but only a single decision of global fit should be made. I propose to combine the multiple EPC tests as follows. First, if any test suggests that the test is underpowered, the test of global fit will be also underpowered due to insufficient sample size. Second, if any test suggests a parameter is severely misspecified, then the model is severely misspecified. If all tests suggest that fixed parameters are trivially misspecified, then the model is trivially misspecified. Otherwise, the decision is inconclusive and reserachers need a larger sample size to get a conclusive result.

3.3 Guidelines for Specifying Parsimony Error

The major problem of the alternative methods and the unified method is that specifying parsimony error is subjective. I think that the subjectivity is necessary as different substantive areas have different views regarding the magnitude of misfit (Steiger, 2004). A misspecified cross loading of .20 may be considered as trivial in one area but meaningful in the other.

Thus, the criterion for acceptable degrees of parsimony error should be developed for each substantive area. To gain an idea of how trivial or severe misspecification are defined by experts in the past research, I reviewed the parameter values used in misspecified models in past simulation studies as well as guidelines in interpreting the values of misspecified parameters. I focused on four types of parameters in the review: standardized cross loadings, factor correlation, residual correlation, and standardized regression coefficients. Because some simulation studies only provided raw parameter values, these values are standardized to facilitate comparison among different studies.

Table 3.1 shows the values of standardized cross loadings for misspecified models across simulation studies. The cross loadings in some studies were below .20 in standardized scale. This value contradicts the guideline used in interpreting standardized loadings (especially in exploratory factor analysis). A higher value, such as .30 (Hair et al., 2006) or .40 (Saris et al., 2009), is considered as the cutoffs between nontrivial and trivial cross loadings. Furthermore, in some studies, the evaluation of standardized cross loadings depends on the size of standardized loadings of target parameters. For example, Beauducel & Wittmann (2005) argued that, if a cross loading was as high as the half of the values of target loadings, then the cross loading was nontrivial. Beauducel & Wittmann (2005) used .4 as the lowest standardized loadings for target parameters so a cross loading of .20 or higher was deemed nontrivial for their designs. Based on their criterion, the standardized loadings of .10 or lower could be considered trivial and the values of .40 or higher could be considered nontrivial.

The values of factor correlations from misspecified models across simulation studies are shown

Table 3.1: The Magnitude of Misspecified Standardized Factor Loadings Reported in Past Simulation Studies on Model Evaluation in SEM.

Studies	Number of Loadings			
	1	2	3	More than 3[a]
Hu & Bentler (1999)	.510	.510, .489		
Beauducel & Wittmann (2005)				(4) .180s
Curran et al. (2003)	.210	.210s	.210s	
Fan et al. (1999)		.422, .429		
Jackson (2007)				(6) .100s
Nye & Drasgow (2011)				**(6) .043-.179**
Taylor (2008)	**.186**	**.186s**	**.186s**	
Wu (2008)	.389	.389, .404		

Note. The boldface values represent the lowest value in each column. Each cell represents a design in a study (represented by rows). If multiple cells in a row are specified, they represent different values of misspecified parameters in a design. If different designs in a study have the same number of factor loadings with different values of misspecified parameters, the lowest magnitude is provided. [a]The value in a parenthesis represents the number of misspecified factor loadings.

in Table 3.2. The lowest value of misspecified factor correlations is .3. Note that the values in Table 3.2 are examples of severe misspecification which do not indicate the minimal severe misspecification. This lowest value is still much higher than the low level of correlation (.1) from Cohen's (1988; 1992) guideline. Saris et al. (2009) also considered the level of .1 as a minimal severe misspecification. Ferguson (2009), however, suggested a threshold of .2 or .3 for small effects and showed that the Cohen's guideline for small effects in standardized mean differences (.2) and correlations (.1) were inconsistent. From this review, the factor correlation of .1 or lower may be considered trivial and the values of .3 or higher may be considered severe.

Residual correlations can be the correlations among residuals of regression effects or among measurement errors. The summary of misspecifications on residual correlations is shown in Table 3.3. The lowest residual correlations is .08. All other residual correlations are .10. The recommendations from Cohen (1988), Ferguson (2009), and Saris et al (2009) described in the previous paragraph are also applicable here because they provide the guidelines for any types of correlations. From this review, the residual correlation of .1 or lower may be considered trivial and the

Table 3.2: The Magnitude of Misspecified Factor Correlation Values from Different Designs across Simulation Studies on Model Evaluation in SEM.

Studies	Number of Factor Correlations			
	1	2	3	More than 3[a]
Davey (2005)	**.400**			
Fan & Sivo (2007)	.854	.854, .859		
Heene et al. (2011)			**.300, .400, .500**	
Hu & Bentler (1999)	.500	**.400, .500**		

Note. The boldface values represent the lowest value in each column. Each cell represents each design in a study (represented by rows). If multiple cells in a row are specified, they represent different values of misspecified parameters in a design. If different designs in a study have the same number of factor correlation with different values of misspecified parameters, the lowest magnitude is provided. a[a]The value in a parenthesis represents the number of misspecified factor correlations.

values of .3 or higher may be considered severe.

Finally, ignoring standardized regression coefficients can be another source of model misspecification. Table 3.4 showed the misspecified values of standardized regression coefficients. All values are higher than or equal to .22. Saris et al. (2009) proposed that the value of .10 or higher is nontrivial. Cohen (1988) did not provide guideline for standardized regression but the thresholds for standardized regression coefficients can be deduced from those for correlation coefficients. For a simple regression analysis, the standardized regression coefficient of Y on X is equal to the correlation between X and Y. Thus the thresholds for correlation coefficiens apply to standardized regression coefficients. Unfortunately, this guideline may be not applicable for partial standardized regression coefficients in multiple regression analysis. Furthermore, Ferguson (2009) indicated that standardized regression coefficients of .2 or .3 are deemed small effects. From this review, the standardized regression of .1 or lower can be considered trivial and the values of .3 or higher are considered severe.

Here I provided a general guideline for researchers in different areas to start thinking about the meaning of trivial misspecifications in their own fields. The maximal trivial parameter values for

Table 3.3: The Magnitude of Misspecified Residual Correlation Values from Different Designs across Simulation Studies on Model Evaluation in SEM.

Studies	Number of Residual Correlations			
	1	2	3	More than 3[a]
Hancock & Mueller (2011)	.405			
Heene et al. (2012)			**.101, .251, .236**	**(6) .080-.236**
Saris et al. (2009)	**.219**			

Note. The boldface values represent the lowest value in each column. Each cell represents each design in a study (represented by rows). If multiple cells in a row are specified, they represent different values of misspecified parameters in a design. If different designs in a study have the same number of residual correlation with different values of misspecified parameters, the lowest magnitude is provided. [a]The value in a parenthesis represents the number of misspecified residual correlations.

Table 3.4: The Magnitude of Misspecified Standardized Regression Values from Different Designs across Simulation Studies on Model Evaluation in SEM.

Studies	Number of Standardized Regression Coefficients			
	1	2	3	More than 3[a]
Curran et al. (2003)				**(4) .350-.450**
Fan & Sivo (2007)	.360			
Heene et al. (2012)	**.220**			
Schermelleh-Engel et al. (2003)	.311	**.311, .570**		

Note. The boldface values represent the lowest value in each column. Each cell represents each design in a study (represented by rows). If multiple cells in a row are specified, they represent different conditions in a study. If different designs in a study have the same number of standardized regression with different values of misspecified parameters, the lowest magnitude is provided. [a]The value in a parenthesis represents the number of misspecified regression coefficients.

each parameter need to be identified for each area of study. For example, researchers may use the guidelines above to pick a value between the trivial and severe misspecification to represent the meaning of trivial misspecification in their fields of study. Furthermore, in setting trivial misspecification, the change in parameter values of target parameters should be investigated. If parameters that are deemed trivial significantly change the parameter values of target parameters (Marsh & Hau, 1996), the parameters are probably not really trivial. The significant change in parameter values may be deduced from the magnitude of the change. If the magnitude of the change alter the qualitative meaning of the parameter, the change is significant. For example, a factor correlation changes from .1 (small effect size) to .3 (medium effect size). For an independent clustered two-factor CFA model, if researchers specify all possible cross loadings equal to .3, the model might fit the data perfectlly well. However, the standardized loadings of target parameters and factor correlations will be highly overestimated (Savalei, 2012). In the random sampling method, researchers can check the estimated target parameter values at each draw and investigate whether their values are highly influenced by the trivial misspecifications.

3.4 Numerical Illustration

I use the Holzinger and Swineford's (1939) data to illustrate the unified approach of model fit. I choose this example because readers may compare the result from Muthén & Asparouhov (2012) on the Bayesian approach, which accounts for trivial misspecification as well. I use the data from the Pasteur school (N = 156). A four-factor CFA model suggested by Muthén & Asparouhov (2012) is used as the target model. Initially, the data were analyzed by MLE. Table 3.5 shows the estimated standardized loadings and factor correlations. The fit indices were as follows: χ^2 (146) = 259.75, $p < .001$, RMSEA = .071 (90% confidence interval: .056 to.085), SRMR = .084, CFI = .884, and TLI = .864.

Table 3.5: Holzinger and Swineford's 1939 Results from Four-Factor Confirmatory Factor Analysis based on Maximum Likelihood Estimation.

Test	Spatial	Verbal	Speed	Memory
Standardized Factor Loadings				
Visual perception	**.814**	.000	.000	.000
Cubes	**.392**	.000	.000	.000
Paper from board	**.441**	.000	.000	.000
Flags	**.578**	.000	.000	.000
General information	.000	**.836**	.000	.000
Paragraph comprehension	.000	**.806**	.000	.000
Sentence completion	.000	**.870**	.000	.000
Word classification	.000	**.724**	.000	.000
Word meaning	.000	**.838**	.000	.000
Addition	.000	.000	**.557**	.000
Code	.000	.000	**.800**	.000
Counting groups of dots	.000	.000	**.517**	.000
Straight and curved capitals	.000	.000	**.537**	.000
Word recognition	.000	.000	.000	**.628**
Number recognition	.000	.000	.000	**.526**
Figure recognition	.000	.000	.000	**.596**
Object-number	.000	.000	.000	**.561**
Number-figure	.000	.000	.000	**.472**
Figure-word	.000	.000	.000	**.467**
Factor Correlations				
Spatial	–			
Verbal	.456	–		
Speed	.362	.485	–	
Memory	.365	.180	.471	–

Based on the results of fit indices, some researchers may reject the model because some fit indices indicated bad fit based on suggested cutoffs (i.e., RMSEA, CFI and TLI based on the Hu and Bentler's guidelines). As mentioned above, the suggested cutoffs should not be used because fit indices are sensitive to many factors, especially model characteristics. I will illustrate the unified approach accounting for trivial misspecification. In this model, some factor loadings and all measurement error correlations are fixed as 0 so these parameters are the potential sources of misspecification. There may be the fifth factor that researchers do not include in a model so I use the ignored factor as another potential source of misspecification. The maximal trivial misspecification levels for all fixed parameters were then defined as (a) standardized cross loadings of .2, (b) measurement error correlation of .2, and (c) ignoring one trivial factor (intentionally for the sake of model parsimony). The magnitude of standardized loadings of the ignored factor was below .2. The factor variance was 1. This factor was not correlated with other target factors in the model. Maximal trivial misspecification and minimal severe misspecification were searched based on the defined parsimony error.

The repeated sampling method was used to find the maximal trivial misspecifications for each fit index. There are 57 cross loadings for homogeneous factor structure with 19 variables and 4 factors. The maximal trivial misspecification was searched by randomly drawing values for the 57 cross loadings from a uniform distribution from -.2 to .2. For 171 measurement error correlations, the uniform distributions from -.2 to .2 were used to draw values. For the ignored factor, three out of 19 loadings were randomly selected to be the nonzero loadings of the factor. Then these three values of loadings were drawn from the uniform distributions of -.2 to .2. One-thousand combinations were drawn and the combinations with the highest values of RMSEA, SRMR, CFI, and TLI were picked as the maximal trivial misspecifications for each fit index. The population misfit of the maximal trivial misspecifications defined by each fit index are shown in Table 3.6. The same combination of misfit was picked as the maximal trivial misspecification defined by RMSEA, CFI, and TLI while SRMR picked a different combination. The misspecified parameter values of both combinations are shown in Appendix B.

Table 3.6: Population Fit Indices of Each Type of Misspecifications based on Sample Size of 156.

Fit Indices	Observed	Maximal Trivial Misspecification			Minimal Severe Misspecification			Decision
		Pop. Fit	Crit. Value	p-value	Pop. Fit	Crit. Value	p-value	
RMSEA	.071	.184	.193	1.000	.002	.000	1.000	Inconclusive
SRMR	.084	.117	.141	1.000	.003	.049	1.000	Inconclusive
CFI	.884	.556	.509	1.000	1.000	1.000	1.000	Inconclusive
TLI	.864	.480	.427	1.000	1.000	1.026	1.000	Inconclusive

Note. RMSEA = Root mean square error of approximation. SRMR = Standardized root mean squared residuals. CFI = Comparative fit index. TLI = Tucker-Lewis index.

The minimal severe misspecification was searched based on the following procedures. First, each combination allowed only either one cross loading, one error correlation, or one omitted factor. The combination with one cross loading had the value of either -.2 or .2. Thus, there were 57 (possible cross loadings) $\times$ 2 (negative or positive) = 104 possible combinations. The combination with one error correlation had the value either -.2 or .2 so there were 171 (possible error correlations) $\times$ 2 (negative or positive) = 342 possible combinations. For the third misspecification, one omitted factor was included. All possible combinations of three omitted loadings from 19 possible loadings were $C_3^{19} = 969$ possible combinations. The loadings of three omitted loadings were (.2, .2, .2), (-.2, .2, .2), (.2, -.2, .2), and (.2, .2, -.2). Thus, there were 969 $\times$ 4 = 3876 possible combinations. Therefore, there were 104 + 171 + 3876 = 4322 combinations to be searched for minimal severe misspecifications. Second, the combinations with the lowest population RMSEA, SRMR, CFI, and TLI were picked as the minimal severe misspecification. For all four fit indices, the minimal severe misspecification was the same combination with an omitted factor with which the factor loadings on Items 1, 3, and 4 are .2. The population misfit values are provided in Table 3.6.

Regarding to the global fit evaluation, the shortcuts (Steps 2a-4a and 5a-7a) can be used. The observed fit indices did not have better fit than the minimal severe misspecifications and worse fit than maximal trivial misspecifications. Therefore, the global fit evaluation result was inconclusive. The direct method based on finding the sampling distributions of fit indices from all maximal trivial misspecifications and minimal severe misspecifications can be used. The critical values from those

sampling distributions are shown in Table 3.6. The observed fit indices did not fall in any critical regions of the tests with maximal trivial misspecifications and minimal severe misspecifications. Therefore, the result was still inconclusive. The plugin *p*-values from those combinations are also shown in Table 3.6, resulting in nonsignificant results. Because all fit indices in global fit evaluation suggested the inconclusive results, the overall global fit evaluation was inconclusive. On the other hand, the Bayesian analysis with the cross-loading normal priors (mean = 0 and variance = 0.01) provided the PPP value of .16. One cross-loading credible interval did not cover 0. The model can be considered as approximate fit with the suggestion for model-fit improvement by freeing the significant cross loading (Muthén & Asparouhov, 2012).

Local fit was evaluated by the confidence intervals of EPC, shown in Appendix C. The confidence intervals were transformed to standardized values to check whether they bracket the maximal trivial misspecification level. The widths of confidence intervals of standardized EPCs were also calculated to check whether the tests were underpowered. The trivial misspecifications of all fixed parameters were from -.2 to .2 in standardized scales. Therefore, if any confidence intervals of standardized EPCs had higher widths than .4, this test would be underpowered. Thirteen confidence intervals of standardized EPCs had the widths greater than .4 so the overall fit was inconclusive. Therefore, the model fit evaluated by the unified approach was inconclusive. A larger sample size should be collected to get a conclusive result.

Chapter 4

Simulation Designs

This dissertation proposes the unified approach for model evaluation. Theoretically, the unified approach should be able to retain trivially misspecified models but reject severely misspecified models. Two simulation studies are used to investigate the performance of the unified approach. The performance of the unified approach is good if the following characteristics are satisfied:

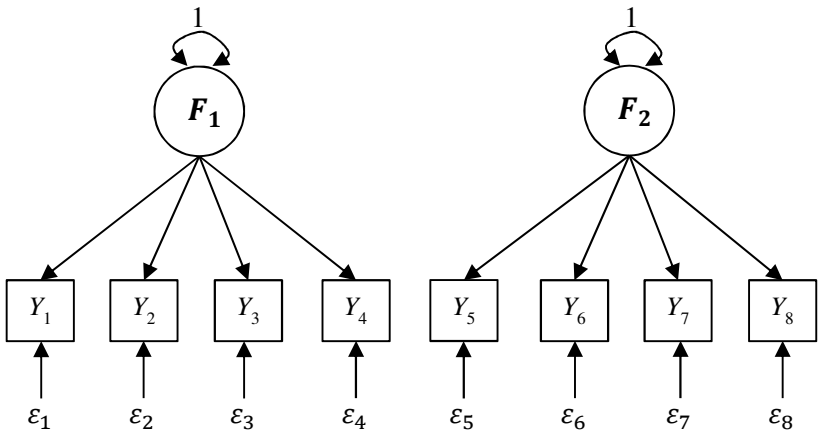
1. The underpowered or inconclusive results should decrease when sample size increases.
2. Among conclusive results, the rejection rate for trivially misspecified models is close to 0.
3. Among conclusive results, the rejection rate for severely misspecified models is close to 1.
4. The performance of the unified approach does not depend on model characteristic such as model size, model type, and incidental parameters.
5. The performance of the unified approach does not depend on sample size.
6. The performance of the unified approach should be consistent across different types of misspecification.

The first simulation aims to compare the performance of the unified approach with other model evaluation methods described in Chapters 1 and 2. The second simulation study evaluates the performance of the unified approach for a different type of models: growth curve models.

4.1 Study 1

The aim of this study is to investigate the performance of the unified approach and to compare the performance of the unified approach with other global fit evaluation methods, including one-size-fit-all fit indices cutoffs, tests of close fit and not close fit, the modification indices and power approach, the Bayesian approach, and the simulation approach. The target model is a CFA model with two uncorrelated factors, as shown in Figure 4.1. The magnitudes of factor loadings are varied across conditions. The measurement error variances are specified in such a way that the total variances of the indicators are equal to 1.

Figure 4.1: The target model for Study 1.

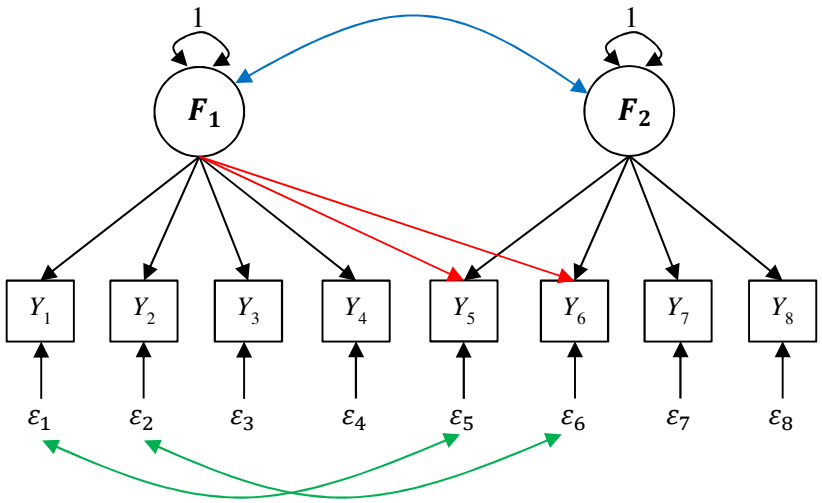


4.1.1 Design Conditions

Data are generated from a model with four degrees of misspecifications varying from perfect fit to very severe misspecification. I refer to these four levels of misspecifications as Levels 0, 1, 2, and 3. Three types of misspecification are considered in this study: (a) omitting the factor correlation, (b) omitting cross loadings, and (c) omitting measurement error correlations (see Figure 4.2). There are 10 combinations of misspecifications as follows:

1. **No Misspecification (Level 0)**: The target model fit the data-generating model perfectly.

Figure 4.2: Types of misspecification for the target model in Study 1. The blue line represents the Type A misspecification. The red lines represent the Type B misspecification. The green lines represent the Type C misspecification.



2. **Type A Misspecification, Level 1**: The target model omits a factor correlation of .1.

3. **Type A Misspecification, Level 2**: The target model omits a factor correlation of .3.

4. **Type A Misspecification, Level 3**: The target model omits a factor correlation of .9.

5. **Type B Misspecification, Level 1**: The target model omits the following standardized factor loadings: $\lambda_{5,1} = .1$ and $\lambda_{6,1} = .1$. Note that the measurement error variances of all indicators in the data-generating models are calculated such a way that the total variances of the indicators are 1 and the target and misspecified factor loadings are standardized.

6. **Type B Misspecification, Level 2**: The target model omits the following standardized factor loadings: $\lambda_{5,1} = .3$ and $\lambda_{6,1} = .3$.

7. **Type B Misspecification, Level 3**: The target model omits the following *unstandardized* factor loadings: $\lambda_{5,1} = 0.9$ and $\lambda_{6,1} = 0.9$. The unstandardized values of target factor load-

ings and measurement error variances remain the same.[1]

8. **Type C Misspecification, Level 1**: The target model omits the following measurement error correlations: $\theta_{1,5} = .1$ and $\theta_{2,6} = -.1$.

9. **Type C Misspecification, Level 2**: The target model omits the following measurement error correlations: $\theta_{1,5} = .3$ and $\theta_{2,6} = -.3$.

10. **Type C Misspecification, Level 3**: The target model omits the following measurement error correlations: $\theta_{1,5} = .9$ and $\theta_{2,6} = -.9$.

These ten data generating models are analyzed by the hypothesized model shown in Figure 4.1. The next design condition is the two levels of trivial misspecification. The first level of trivial misspecification is to use the cutoffs of .1 (Level 1) as the maximal trivial misspecification for the factor correlation, the standardized loadings, and the measurement error correlations. In this case, Level 0 degree of misspecification is considered trivial and Levels 2 and 3 degrees of misspecification are considered severe. Level 1 degree of misspecification is equal to the level of maximal trivial misspecification. This condition is referred to as the cutoff conditions. It is unclear to consider the cutoff condition as trivial or severe misspecification.[2] The second level of trivial misspecification is the cutoffs of .3 (Level 2). In this case, Levels 0 and 1 degrees of misspecification are considered trivial, Level 2 degree of misspecification is the cutoff condition, and Level 3 degree of misspecification is considered severe. The same value of .1 or .3 is used for the factor correlation, the standardized loadings, and the measurement error correlation as the level of trivial misspecificaiton for the sake of simplicity. The values of .1 and .3 are the minimum and maximum values that are deemed trivial according to some of the guidelines that I have reviewed in Chapter 3.

[1]The unstandardized loadings are used because the specification of standardized values as .9 would provide negative error variances when the total indicator variances are fixed as 1. For example, if the target factor loadings are .5 and the total variance is fixed as 1, the measurement error variance would be $1 - .9^2 - .5^2 = -.06$.
If the target unstandardized factor loadings are 0.5, the standardized loadings would be as follows: $\lambda_{5,1} = .497$, $\lambda_{5,2} = .276$, $\lambda_{6,1} = .497$, and $\lambda_{6,2} = .276$. If the target unstandardized factor loadings are 0.7, the standardized loadings would be as follows: $\lambda_{5,1} = .497$, $\lambda_{5,2} = .387$, $\lambda_{6,1} = .497$, and $\lambda_{6,2} = .387$.

[2]If the degree of misspecification is slightly lower than .1, it is classified as trivial misspecification. On the other hand, if the degree of misspecification is slightly higher than .1, it is classified as severe misspecification.

Regarding the other design conditions, the standardized target factor loadings are .5 or .7. The numbers of indicators are 8 or 16. The details of how the misspecifications are imposed in the model with 16 items are provided in Appendix D. Sample size are 125, 250, 500, 1000, 2000, or 4000. Thus, there are 10 (data generating models) $\times$ 2 (the levels of trivial misspecification) $\times$ 2 (factor loadings) $\times$ 2 (the numbers of items) $\times$ 6 (sample size) $=$ 480 conditions in this simulation study.

4.1.2 Procedures for the Unified Approach

Below I describe the steps to implement the unified approach for the model considered in this study. For global fit evaluation, the unified approach is implemented as follows:

1. The CFA model is fitted on observed data. Parameter estimates and four fit indices (RMSEA, SRMR, CFI, and TLI) are saved.

2. Find the maximal trivial misspecification by the repeated sampling method from the following parameter spaces:

   (a) One factor correlation ranges from -.1 to .1 (or -.3 to .3).

   (b) Standardized cross loadings range from -.1 to .1 (or -.3 to .3). The numbers of cross loadings for 8- and 16-indicator models are 8 and 16, respectively.

   (c) Measurement error correlations range from -.1 to .1 (or -.3 to .3). The numbers of measurement error correlations for 8- and 16-indicator models are 28 ($8 \times 7/2 = 28$) and 120 ($16 \times 15/2 = 120$), respectively.

   I use 1,000 draws from the uniform distributions across the ranges of all misspecifications. Population RMSEA, SRMR, CFI, and TLI are computed for each of the 1,000 draws.

3. Each fit index value from observed data is compared with the values from all combinations. If the former indicates a better fit than the latter, the test with maximal trivial misspecification is not significant for the fit index (Step 2a-4a).

4. If the observed fit index indicates a worse fit than all combinations, the combination providing the maximal misfit is used to generate 1,000 data sets. Then, the target model (two-factor CFA model) is fit to all generated data. If the observed fit index falls in the critical region, which is the area covering 5% of the poor-fit extreme, the target model is severely misspecified regarding to the tested fit index (reject the null hypothesis). Otherwise, the target model is either trivially misspecified or severely misspecified (fail to reject the null hypothesis).

5. Find the minimal severe misspecification by listing all possible severe misspecifications in one fixed parameter. All fixed parameters and their changing point are defined as follows:

   (a) One factor correlation is set as either -.1 or .1 (or either -.3 or .3). There are 2 sets for this misspecification.

   (b) Each standardized cross loading is set as either -.1 or .1 (or either -.3 or .3). There are 16 and 32 sets of misspecifications for the 8- and 16-indicator models, respectively.

   (c) Each measurement error correlation is set as either -.1 or .1 (or either -.3 or .3). There are 56 and 240 sets of misspecifications for the 8- and 16-indicator models, respectively.

   The total numbers of misspecification sets are 74 $(2+16+56)$ and 274 $(2+32+240)$ for 8- and 16-indicator models, respectively. Population RMSEA, SRMR, CFI, and TLI are calculated for each combination.

6. Each fit index value from observed data is compared with the values from all listed combinations. If the observed fit index indicates a worse fit than the values from any combinations, the test with minimal severe misspecification is failed to reject (Step 5a-7a).

7. If the observed fit index indicates a better fit than all combinations, the combination providing the minimal misfit is used to generate 1,000 data sets. Then, the hypothesized model (two-factor CFA model) is fit to all generated data. The critical region is the area covering 5% of the good-fit extreme. If the observed fit index falls in the critical region, the hypothesized model is trivially misspecified regarding to the tested fit index (reject the null hypothesis).

Otherwise, the hypothesized model is trivially misspecified or severely misspecified (fail to reject the null hypothesis).

8. If any fit indices (RMSEA, SRMR, TLI, or CFI) are significantly worse fit than the maximal trivial misspecification, the hypothesized model is severely misspecified. If each fit index indicates a significantly better fit than the minimal severe misspecification, the hypothesized model is trivially misspecified. Otherwise, the decision is inconclusive.

9. It is possible that the decisions from the four fit indices are inconsistent. If any fit indices indicate a severe misspecification, the hypothesized model is deemed severely misspecified. If all fit indices indicate a trivial misspecification, the hypothesized model is then deemed trivially misspecified. Otherwise, the overall decision is inconclusive.

The local fit evaluation is implemented regardless of the results from global fit evaluation. The procedure of the local fit evaluation is described as follows:

1. Confidence intervals of EPCs is calculated by Equation 2.1 for each fixed parameter.

2. All fixed parameters have the range of trivial misspecifications as follows:

   (a) The range between -.1 and .1 (or -.3 and .3) is deemed trivial for the factor correlation.

   (b) The range between -.1 and .1 (or -.3 and .3) is deemed trivial for the standardized factor loadings.

   (c) The range between -.1 and .1 (or -.3 and .3) is deemed trivial for the measurement error correlations.

3. The widths of the confidence intervals of EPCs are calculated. The confidence intervals that their widths are larger than the range of trivial misspecifications (.2 or .6 in this study) are underpowered.

4. If both lower and upper bounds of the confidence intervals fall within the range of trivial misspecifications, the fixed parameters are trivially misspecified.

5. If both lower and upper bounds of the confidence intervals fall outside the range of trivial misspecifications, the fixed parameters are severely misspecified.

6. If the confidence intervals are partially overlapped with the range of trivial misspecification, the decisions are inconclusive such that the fixed parameters can be either severely or trivially misspecified.

7. The decisions from all fixed parameters are combined to make the inference about global fit. If any of the confidence intervals is underpowered, the model fit evaluation is underpowered. If any of the confidence intervals is severely misspecified, the target model is severely misspecified. If all confidence intervals are trivially misspecified, the target model is trivially misspecified. Otherwise, the decision is inconclusive.

The unified approach is implemented by the `lavaan` package (Rosseel, 2012) in the `R` environment (R Development Core Team, 2013).

4.1.3 Procedures for Other Methods

Five alternative model evaluation methods are compared with the unified approach in the study. The methods are implemented by the `lavaan` package (Rosseel, 2012) in the `R` environment (R Development Core Team, 2013) except that the Bayesian approach is implemented by `Mplus` (Muthén & Muthén, 2013).

4.1.3.1 One-size-fit-all Cutoffs

I use the cutoffs suggested by Hu & Bentler (1999) for the four fit indices: RMSEA < .06, CFI > .95, TLI > .95, and SRMR < .09. If observed fit indices satisfy these inequalities, the hypothesized model is retained. Otherwise, the hypothesized model is rejected. The results from RMSEA, CFI, TLI, and SRMR are individually reported. Furthermore, the information from all four cutoffs is combined such that a model is retained if all cutoffs are satisfied. I acknowledge that Hu and

Bentler's cutoffs were not designed for the model considered in the study. I use them considering the cutoffs are widely used in practice.

4.1.3.2 Test of close fit and not close fit

I use the population RMSEA of .05 as the maximal trivial misspecification (Browne & Cudeck, 1992). If the test of not close fit ($H_0 : \varepsilon \geq .05$) is rejected, the hypothesized model is deemed trivially misspecified. If the test of close fit ($H_0 : \varepsilon \leq .05$) is rejected, the hypothesized model is deemed severely misspecified. If neither hypotheses is rejected, the decision is inconclusive. Note that only RMSEA is used in the study because theoretical sampling distribution cannot be derived for the other fit indices.

4.1.3.3 Modification indices and power approach

I follow the original Saris and colleague's (2009) guideline for model evaluation (see Chapter 2). For each fixed parameter, the result can be either a severe misspecification, a trivial misspecification, or inconclusive. Saris and colleague's (2009), originally, provided a method to search for the sources of misspecification in a model but they did not provide a method to combine sources of misspecification for global fit evaluation. Thus, I use the method similar to the local fit evaluation in the unified approach to combine information from different sources of misspecification. The hypothesized model is rejected if at least one fixed parameter indicates a severe misspecification. If the results from all fixed parameters indicate a trivial misspecification, the hypothesized model is deemed trivially misspecified. Otherwise, the decision is inconclusive.

4.1.3.4 Bayesian approach

Two methods of model evaluation are used in Bayesian approach. First, PPP is only used in model evaluation. The hypothesized model will be rejected if PPP is less than .05 (Muthén & Asparouhov, 2012). Second, PPP is combined with the zero coverage of the credible intervals of nontarget parameters. If PPP is less than .05 or any credible intervals of nontarget parameters do not include

0, the hypothesized model is deemed severely misspecified. Otherwise, the hypothesized model will be retained.

Noninformative priors are used for all target parameters. Two sets of informative priors are used for nontarget parameters. In the first set, the nontarget factor correlation and all cross loadings have normal priors with the mean of 0 and the variance of $(0.1/2)^2 = 0.0025$ (95% confidence limit = ± 0.1) or $(0.3/2)^2 = 0.0225$ (95% confidence limit = ± 0.3), depending on the level of trivial misspecification. In the second set, the nontarget factor correlation remains the same as the first set. The error covariances have normal priors with the mean of 0. The variance of the priors were calculated by $(m \times (1 - l^2)/2)^2$, where m is the level of trivial misspecification in error correlation (0.01 or 0.03, which create 95% confidence limit of ± 0.1 or ± 0.3) and l is the standardized value of target factor loading (0.5 or 0.7). Multiple sets of priors will be used because the Bayesian approach is likely to be influenced by different choices of priors. In sum, there are four results from the Bayesian analysis based on two methods of model evaluation and two sets of informative priors. I use 90,000 burn-in iterations and collecting the next 10,000 iterations for estimating posterior distributions and PPP. The sample is thinned by using every 10 iterations. The Bayesian analysis is considered convergent if the potential reduction scale factor at the 90,000-th iteration is less than 1.1.

4.1.3.5 Simulation approach

Three alternative models are used to represent maximal trivial misspecifications:

1. The fixed factor correlation is set as 0.1 (or 0.3).

2. Two standardized cross factor loadings are set as follows: $\lambda_{5,1} = .1$ (or .3) and $\lambda_{6,1} = .1$ (or .3) for the 8-indicator model (see Appendix D for the details for the 16-indicator model).

3. Two measurement error correlations are changed: $\theta_{1,5} = .1$ (or .3) and $\theta_{2,6} = -.1$ (or -.3) for the 8-indicator model (see Appendix D for the details for the 16-indicator model).

A plugin p-value is derived from the sampling distribution of a fit index established by fitting the hypothesized model to the simulated data sets from trivially misspecified models. If $p \leq .05$, the hypothesized model is rejected. If $p > .05$, the hypothesized model is retained. The results from three maximal trivial misspecifications are also combined. The hypothesized model is rejected if at least one result indicates a severe misspecification. If all results indicate trivial misspecification, the hypothesized model is deemed trivially misspecified.

4.1.4 Simulation Analysis

The one-size-fit-all cutoffs, the Bayesian approach, and the simulation approach provide only two possible outcomes: trivial or severe misspecifications. I refer to these methods as two-outcome methods. The unified approach, the tests of close and not close fit approach and the modification indices and power approach provide three outcomes: trivial misspecifications, severe misspecifications, or inconclusive. I refer to these methods as three-outcome methods. For the three-outcome methods, two outcome variables are obtained: the proportion of inconclusive (or underpowered) results and the proportion of model rejection among conclusive results. If the proportion of inconclusive results is greater than .90, the rejection rate would be computed based on a small proportion of conclusive results so I code the rejection rate as missing in this case. On the other hand, two-outcome methods will provide only the proportion of model rejection.

I mainly use two tables to evaluate the influence of the design conditions on the performance of the model evaluation methods. These tables present results from factorial analysis of variance (ANOVA). The design factors in this analysis include the size of target loadings, the number of indicators, sample size, the degree of misspecifications, the level of maximal trivial misspecification, and the type of misspecifications. However, all factors are not fully crossed. Type of misspecification is only applicable when the level of misspecification is not Level 0. Therefore, in the factorial ANOVA, I did not include the results with Level 0 degree of misspecification. The dependent variables for the first and the second tables are the rejection rates and the proportions of inconclusive results, respectively. The proportions of inconclusive results are only applicable for the three-

outcome methods. Note that each combination of the design factors has only one observation of the dependent variables. Therefore, the highest-order factor cannot be separated from the error variance. The factors with eta-squares (η^2s) of .03 or higher are deemed non-negligible factors. The past research has used .01 (Lüdtke et al., 2008) or .05 (Geldhof et al., in press) as threshold for meaningful η^2s. I found that, in the current study, .01 included factors with negligible effects whereas .05 ignored important effects. Thus, I chose .03 as the cutoff.

When any main or interaction effect has a non-negligible effect, I describe the pattern of the effect. In addition, the averages of the dependent variable across each level of the non-negligible conditions are tabulated to further explain the effect. If an effect is related to the degree of misspecification, the results associated with Level 0 degree of misspecification are included. Given that the interaction between the degree of misspecification and the level of trivial misspecification is the primary research question of the study. The tables of the averages of the rejction rates and the proportion of inconclusive results across the degree of misspecification and the level of trivial misspecification are provided even if the effects are negligible. If the interaction is dependent on other factors (i.e., three-way or higher interaction), the table describing the higher order interaction is shown instead. All tables mentioned above are used to answer the research questions as follows.

4.1.4.1 The Comparisons between Model Evaluation Methods

All model evaluation methods are evaluated in terms of to what extent they have satisfied the desired properties. The desired properties are as follows:

Appropriate Rejection Rates for Varying Degrees of Misspecification and Levels of Trivial Misspecification. The hypothesized model should be rejected if the model misspecification is higher than the maximal level of trivial misspecification. In contrast, the hypothesized model should be retained if the misspecification is lower than the maximal level of trivial misspecification.

From the tables described above, four results are used to indicate whether this desired characteristic is satisfied. First, the degree of misspecification and the level of maximal trivial misspecification should interactively influence the rejection rate. That is, a good model evaluation method

should have high η^2 on the interaction between the degree of misspecification and the level of maximal trivial misspecification. Second, the rejection rates of trivial misspecification conditions should be close to 0. I consider the rejection rates lower than .10 as desirable. Third, the rejection rates of severe misspecification conditions should be close to 1. I consider the rejection rates higher than .90 as desirable. Fourth, the rejection rates of the cutoff conditions (the model misspecification is at the same level as the maximal trivial misspecification) should be in between 0 and 1 because it is unclear to be considered the cutoff conditions as trivial or severe misspecification. I consider the rejection rates between .10 and .90 as desirable.

Rejection Rates Are Not Influenced by Types of Misspecification. There are three types of misspecification in this simulation: the misspecification in factor correlation, the misspecification in cross loadings, and the misspecification in error correlations. A good model evaluation method should be able to detect all types of severe misspecifications. Thus, a good model evaluation method should have low η^2s ($<$ .03) on the main and interaction effects involving the type of misspecification.

Rejection Rates Are Not Influenced by Model Characteristics. This simulation investigates two model characteristics: the number of items (8 and 16) and the magnitude of target factor loadings (0.5 and 0.7). The rejection rates should be consistent across the numbers of items and the magnitudes of target factor loadings. Thus, a good model evaluation method should have low η^2s ($<$ .03) for the main and interaction effects involving the number of items or target factor loadings.

Rejection Rates Are Not Influenced by Sample Sizes. A good model evaluation method should consistently retain trivially misspecified models and reject severely misspecified models across all sample sizes (125, 250, 500, 1000, 2000, and 4000). The effect of sample sizes on the proportion of inconsistent results is investigated in the next section. A good model evaluation method should have low η^2s ($<$ .03) for the main and interaction effects involving sample sizes.

4.1.4.2 The Properties of the Unified Approach

Because the unified approach provides three possible outcomes and consists of both global and local fit evaluations, studying the pattern of the proportion of inconclusive results and the congruency between the global and local model evaluations would provide a better understanding of the performance of the approach.

Pattern of the Proportions of Inconclusive Results. The main benefit of the unified approach is that it does not provide the decision of model rejection (reject vs. retain a model) if it does not have enough information. There are two situations indicating low information:

1. Sample size is low.
2. The degree of misspecification is close to the level of maximal trivial misspecification (especially in the cutoff conditions).

Therefore, the η^2s on the proportion of inconclusive results should be high for the main effect of sample size and the interaction effect between the degree of misspecification and the level of maximal trivial misspecification. Furthermore, the table of the average proportion of inconclusive results classified by the degree of misspecification and the level of maximal trivial misspecification will be shown. The proportion of inconclusive results is expected to be the highest when the sample size is smallest holding the other factors constant and at the cutoff condition holding sample size constant.

The Congruency between Global and Local Model Evaluation. The global model evaluation can provide three possible outcomes: trivial, severe, or inconclusive. The local model evaluation can provide four possible outcomes: trivial, severe, inconclusive, or underpowered. I use a contingency table to see the interaction between the results from both methods. The contingency table is used to see whether global or local fit evaluation is unnecessary to be used. One method is unnecessary if all information from the method is already provided by the other method. For example, global fit evaluation would be unnecessary if the following conditions are satisfied. (a) When the global evaluation provides trivial or severe outcomes, the local evaluation provides the

same outcomes. (b) When the global evaluation provides inconclusive outcomes, the local evaluation provides trivial, severe, or inconclusive outcomes. In this case, the information from the global fit evaluation is covered by the information provided by the local fit evaluation; thus, it is not necessary to be used. If both methods are necessary, I will investigate the conditions under which global or local evaluations have a higher power to detect trivial or severe misspecifications. The contingency table is also used to examine the mismatches between both methods. I check the situation where one method indicates severe misspecification while the other indicates trivial misspecification. This situation should not occur.

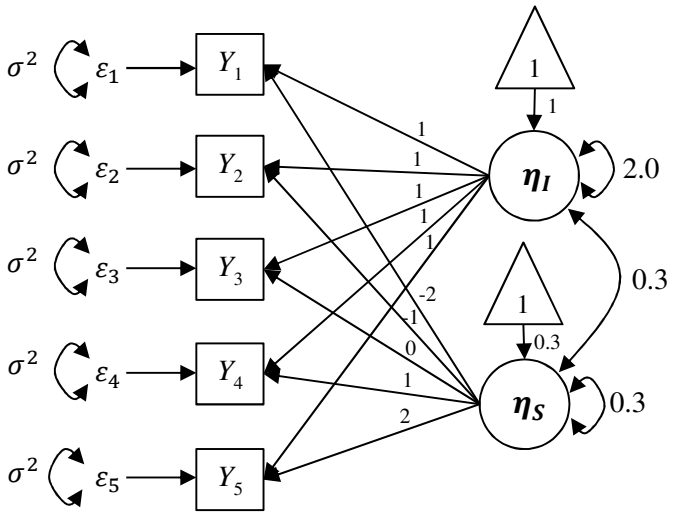
4.2 Study 2

The primary purpose of this study is to investigate the performance of the unified approach for growth curve models. Growth curve models have two sources of model misspecifications: mean and covariance structures. Population practical fit indices, such as RMSEA, SRMR, or CFI, were designed to primarily detect the misfits in the covariance structure so they are not sensitive to the misspecification in the mean structure (Wu & West, 2010). Wu & West (2013) showed that four correlation-based fit indices were able to detect the misfit at the mean structure. The cutoffs, however, were not established so analysts cannot decide when to reject or retain their models based on these correlation-based fit indices. This study examines whether the unified approach is able to correctly classify between trivial and severe misspecifications in mean and covariance structures.

The growth curve model in this study is shown in Figure 4.3. A variable is measured at five time points with linear trajectory. The error variances are constant across time. The intercept variance is 2. The linear slope variance is 0.3. The covariance between intercept and linear factors is 0.3. The means of the intercept and linear slope are 1 and 0.3, respectively.

The matrix presentation of the model is shown here because these notations will be used for specifying trivial misspecifications and calculating correlation-based fit indices described later. Let $\boldsymbol{y}_i$ be the observed score vector of Participant i with the length of 5 (the number of time points).

Figure 4.3: The target model for Study 3.



The growth curve model described in Figure 4.3 expresses the observed score vector as follows:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{4.1}$$

where $\boldsymbol{\eta}_i$ is the latent variable vector of Participant i with the length of 2. The first and second elements of the vector represents the latent variable scores of intercept (i.e., the expected value of marginal mean at Time 3) and linear change. $\boldsymbol{\eta}_i \sim MVN(\boldsymbol{\alpha}, \boldsymbol{\Psi})$ where MVN represents a multivariate normal distribution, $\boldsymbol{\alpha}$ is the latent variable means across participants, and $\boldsymbol{\Psi}$ represents a 2×2 covariance matrix among latent variables. $\boldsymbol{\varepsilon}_i$ represents errors of Participant i. $\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}_5, \boldsymbol{\Theta})$ where $\mathbf{0}_5$ is the zero vector with the length of 5 and $\boldsymbol{\Theta}$ represents a 5×5 covariance matrix among residuals.

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}. \tag{4.2}$$

The model-implied means ($\hat{\boldsymbol{\mu}}$) and covariance matrix ($\hat{\boldsymbol{\Sigma}}$) can be calculated as follows:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}, \tag{4.3}$$

and

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\hat{\boldsymbol{\Psi}}\boldsymbol{\Lambda}' + \hat{\boldsymbol{\Theta}}. \tag{4.4}$$

The estimated individual-specific growth parameters ($\hat{\boldsymbol{\eta}}_i$) represents the estimated intercept and linear change of Participant i combining information from the observed scores ($\boldsymbol{y}_i$) and the average of latent variable scores ($\boldsymbol{\alpha}$):

$$\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Psi}}\boldsymbol{\Lambda}'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\alpha}}. \tag{4.5}$$

The predicted values of individual responses or conditional means of Participant i ($\hat{\boldsymbol{y}}_i$) can be calculated:

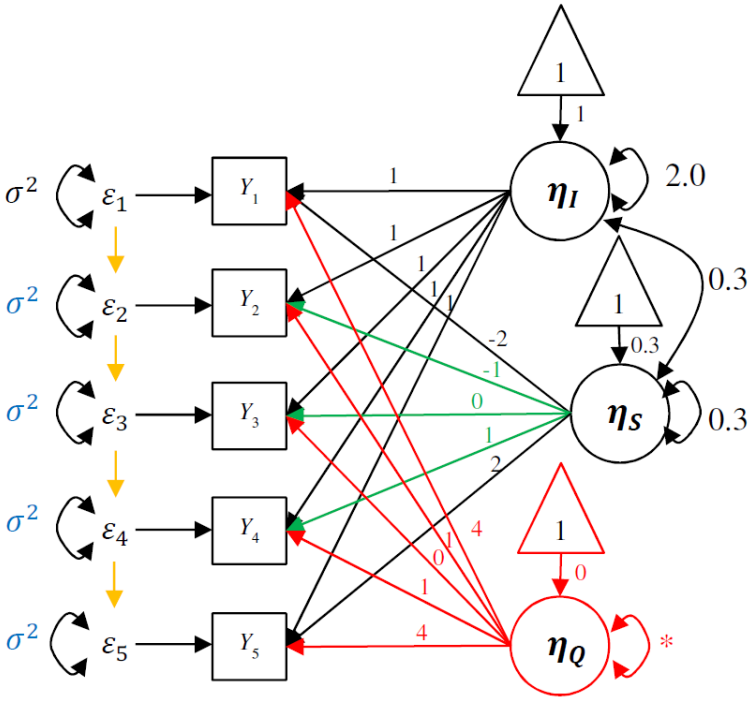
$$\hat{\boldsymbol{y}}_i = \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}_i. \tag{4.6}$$

4.2.1 Design Conditions

The design conditions are similar to Study 1. Data are generated from the model with four degrees of misspecification ranged from perfit fit to very severe misspecification. I refer these four levels of misspecifications as Levels 0, 1, 2, and 3. Four types of misspecification are considered in this study: (a) omitting first-order autoregressive regression among residuals, (b) omitting quadratic factor, (c) nonlinear change, and (d) unequal residual variances (see Figure 4.4). There are 13 combinations of misspecifications as follows:

1. **No Misspecification or Level 0**: The target model fits the data-generating model perfectly.

Figure 4.4: Types of misspecification for the target model in Study 3. The orange lines represent the Type A misspecification. The red lines represent the Type B misspecification. The green lines represent the Type C misspecification. The blue texts represent the Type D misspecification.



2. **Type A Misspecification, Level 1**: The target model omits the first-order autoregressive regression among residuals. The standardized regression coefficient is .1.

3. **Type A Misspecification, Level 2**: The target model omits the first-order autoregressive regression among residuals. The standardized regression coefficient is .3.

4. **Type A Misspecification, Level 3**: The target model omits the first-order autoregressive regression among residuals. The standardized regression coefficient is .7.

5. **Type B Misspecification, Level 1**: The target model omits the quadratic factor. The mean and variance of the quadratic factor are 0 and 0.003364, respectively.

6. **Type B Misspecification, Level 2**: The target model omits the quadratic factor. The mean and variance of the quadratic factor are 0 and 0.016384, respectively.

7. **Type B Misspecification, Level 3**: The target model omits the quadratic factor. The mean and variance of the quadratic factor are 0 and 0.065025, respectively.

8. **Type C Misspecification, Level 1**: The change of the population model is not linear. Rather, the factor loadings of the so-called linear factor on all indicators are -2, -0.67, 0.37, 1.98, and 2 (see the blue solid line in Figure 4.5).

9. **Type C Misspecification, Level 2**: The change of the population model is not linear. Rather, the factor loadings of the so-called linear factor on all indicators are -2, -0.22, 0.88, 3.31, and 2 (see the green dashed line in Figure 4.5).

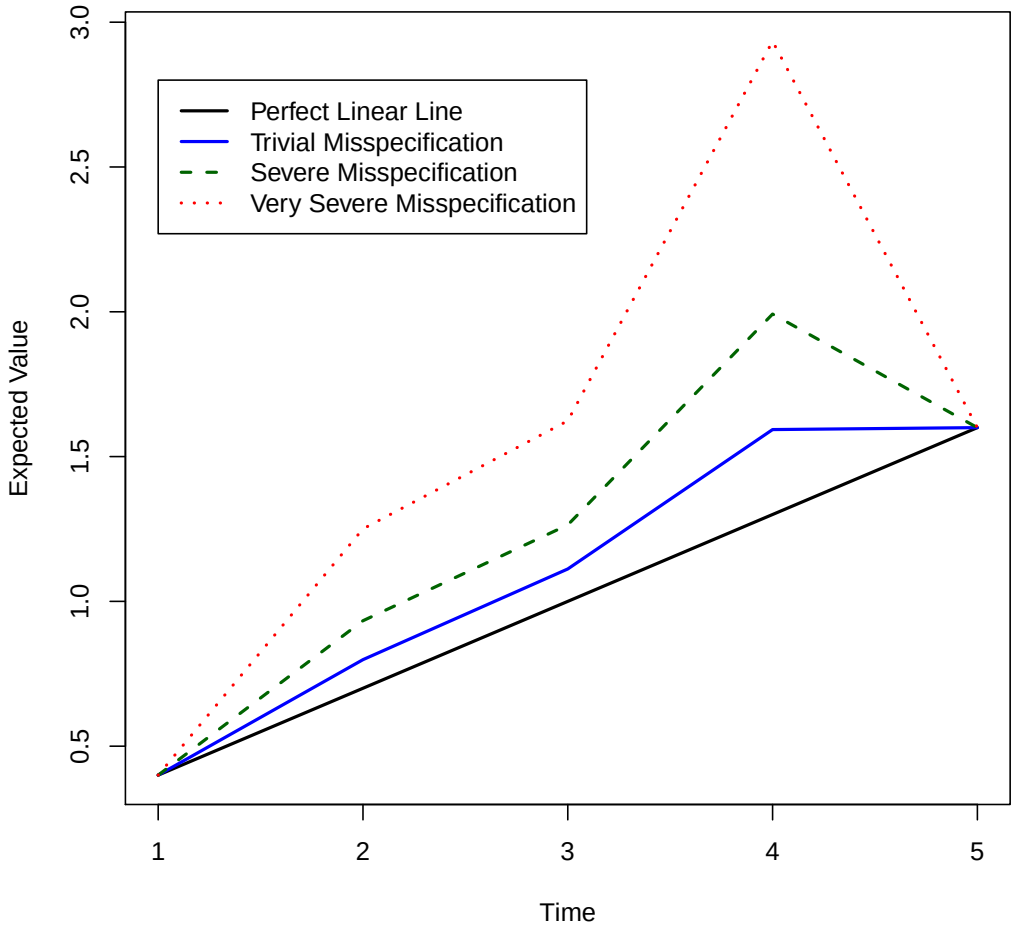
10. **Type C Misspecification, Level 3**: The change of the population model is not linear. Rather, the factor loadings of the so-called linear factor on all indicators are -2, 0.83, 2.08, 6.45, and 2 (see the red dotted line in Figure 4.5).

11. **Type D Misspecification, Level 1**: The residual variances linearly increase over time. The residual variance of the fifth time points is 1.376 times of the residual variance of the first time point.

12. **Type D Misspecification, Level 2**: The residual variances linearly increase over time. The residual variance of the fifth time points is 2.78 times of the residual variance of the first time point.

13. **Type D Misspecification, Level 3**: The residual variances linearly increase over time. The residual variance of the fifth time points is 23.84 times of the residual variance of the first time point.

Figure 4.5: The expected means at each time point with the Type C misspecification imposed.



These thirteen data generating models are analyzed by the hypothesized model shown in Figure 4.3. The degree of misspecification is designed such that Levels 0, 1, 2, and 3 are no, trivial, severe, and very severe misspecification, respectively. The level of maximal trivial misspecification is in between Levels 2 and 3, which will be described in the next section. Note that standardized regression coefficients (Type A misspecification) of .1, .3, and .7 are considered as trivial, severe, and very severe misspecification based on the review in Chapter 3. For Type C misspecification, I assume that two trends (i.e., the change in means over time) are trivially different if the proportion of variance of one trend explained by the other trend is .90. This value is based on the suggestion

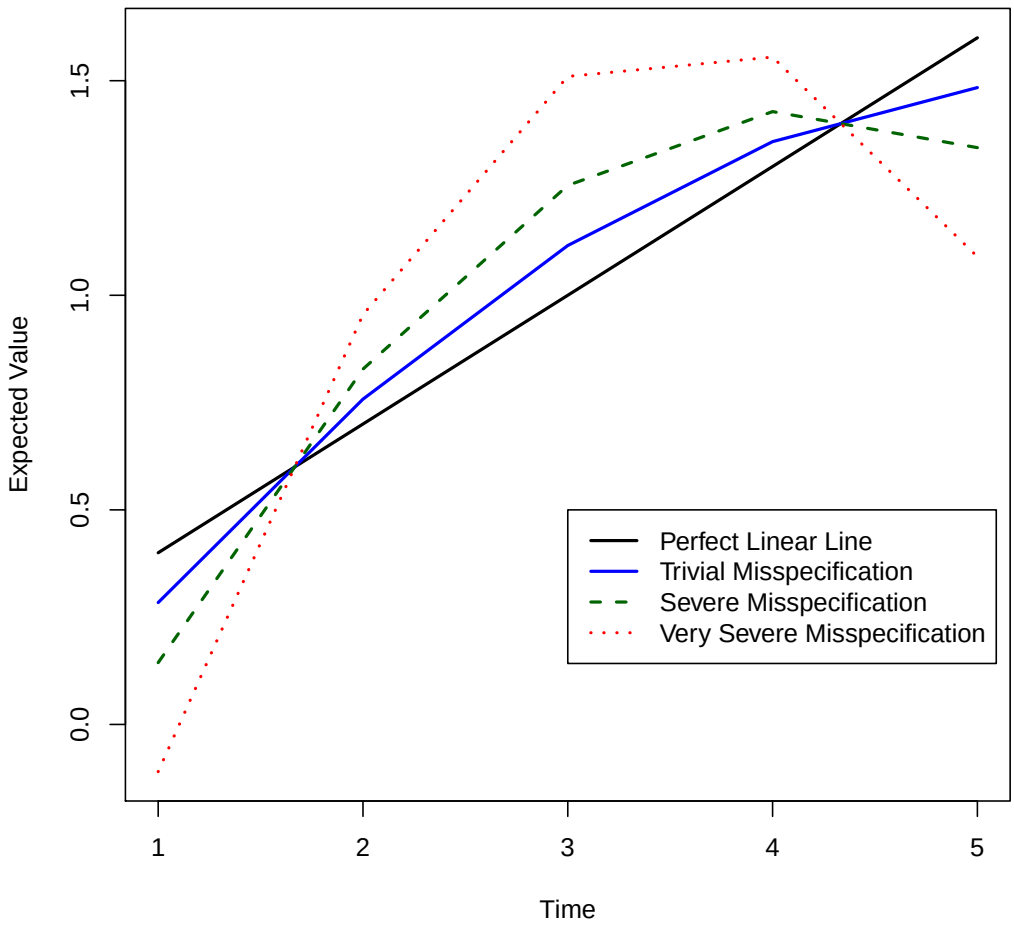
of the multicollinearity problem in multiple regression (Tabachnick & Fidell, 2012). The trivial, severe, and very severe Type C misspecifications have a proportion of variance explained by the perfect linear trend of .95, .80, and .50, respectively.

Regarding to Type B misspecification, I will calculate the model-implied means ($\hat{\boldsymbol{\mu}}$) when the latent variable mean of the quadratic trend is specified as one standard deviation below the mean and the means of the intercept and the linear slope are specified as 0.[3] These model-implied means for trivial, severe, and very severe misspecification are shown in Figure 4.6. Then, the model-implied means are then predicted by the linear trend and the proportion of variance explained by the linear treand is calculated. For trivial, severe, and very severe Type B misspecifications, the proportions of variance explained are .95, .80, and .50, respectively.

Regarding to Type D misspecification, the ratios of the maximum and minimum error variances will be used to quantify the levels of misspecifications. To my knowledge, there is no guideline of the effect sizes of the ratios of variances. Hence I quantify the effect sizes of the ratios of variances by finding the values of the ratios such that they provide the same power as in detecting standardized regression coefficient in a simple regression. As mentioned above, the standardized coefficients of .1, .3, and .7 are deemed trivial, severe, and very severe, respectively. First, I calculate the sample sizes needed to detect these levels of standardized regression coefficients given the statistical power of .80. Then, those sample sizes are used to find the ratio of variances in F test providing statistical power of .80. The ratios of variances of 1.376, 2.78, and 23.84 are equivalent to the standardized regression coefficients of .1, .3, and .7.

[3]The factor loadings of the quadratic factor are (4, 1, 0, 1, 4).

Figure 4.6: The expected means at each time point when the misspecified quadratic factor (Type B misspecification) has the value at one standard deviation below the mean.



Regarding to the other design conditions, the residual error variances of the first time point are 0.5, 2, or 3.5. Sample size will be 125, 250, 500, or 1000. Thus, there will be $13 \times 3 \times 4 = 156$ conditions in this simulation study.

4.2.2 Procedures for the Unified Approach

The process of the unified approach is similar to Study 1 except some model-specific steps. First, the maximal trivial misspecifications are based on the following parameter spaces:

1. Ten residual correlations between observed variables between five time points ($5 \times 4/2 = 10$) range from -.2 to .2

2. Residual variances are different across time points so that the ratio between the maximum and minimum variances is not over 1.926 (equivalent to the standardized regression coefficient of .2). To get the residual variances, initially, five numbers are drawn from the uniform distributions between 1 and 1.926, denoted as c_1, c_2, c_3, c_4, and c_5. Next, the residual variance of the first time point is the estimated common residual variance from data analysis. The residual variances of the second, third, fourth, and fifth time points will be the the residual variance of the first time point multiplied by c_2/c_1, c_3/c_1, c_4/c_1, and c_5/c_1, respectively.

3. The factor loadings of the linear trend ($\boldsymbol{\lambda}_2$) are deviated (denoted as $\tilde{\boldsymbol{\lambda}}_2$) such that the proportion of variance of the deviated marginal means ($\tilde{\boldsymbol{\mu}}$) explained by the original marginal means ($\boldsymbol{\mu}$) is .90 or higher. This proportion of variance is denoted as ρ^2. ρ^2 is randomly drawn from the uniform distribution between 0.9 and 1. $\tilde{\boldsymbol{\lambda}}_2$ is produced by the following procedure:

   (a) Let $\boldsymbol{e}$ be a vector with five elements created from $N(0, \sqrt{1-\rho^2})$. The values are then transformed to make sure that the observed mean and standard deviation of five observations are 0 and $\sqrt{1-\rho^2}$.

   (b) The vector of linear trend, $\boldsymbol{\lambda}_2 = (-2,-1,0,1,2)$ is standardized so that the mean is 0 and the standard deviation is 1. The standardized linear trend is denoted as $\boldsymbol{\lambda}_{2S}$.

   (c) $\boldsymbol{e}$ is regressed on $\boldsymbol{\lambda}_{2S}$. The residuals of the regression are saved and denoted as $\boldsymbol{e}_I$. This step is used to make sure that the residuals are independent from $\boldsymbol{\lambda}_{2S}$.

   (d) The standardized deviated linear trend, $\tilde{\boldsymbol{\lambda}}_{2S}$, is calculated by $\tilde{\boldsymbol{\lambda}}_{2S} = \sqrt{\rho^2} \cdot \boldsymbol{\lambda}_{2S} + \sqrt{1-\rho^2} \cdot \boldsymbol{e}_I$.

   (e) The deviated linear trend is transformed such that the first and the last point are -2 and 2. This transformation is implemented by the following formula: $\left(4 \times (\tilde{\lambda}_{t2} - \tilde{\lambda}_{12})/(\tilde{\lambda}_{52} - \tilde{\lambda}_{12})\right) - 2$.

4. A quadratic trend is omitted. This quadratic trend has the mean of 0. The variance of this quadratic trend is randomly drawn from a uniform distribution from 0 to 0.083 (the expected value of the change when the value of quadratic trend equals one standard deviation has the variance explained by the linear trend of .90 to 1.00). The factor loadings of the quadratic coefficient will be (4, 1, 0, 1, 4). This quadratic factor is independent of other latent variables.

The minimal severe misspecification is searched among all possible minimal severe misspecifications for each dimension of fixed parameters. All fixed parameters and their changing point are defined as follows:

1. Each of the 10 residual correlations among the 5 repeated measures is set as either -.2 or .2. There are 20 sets from this misspecification.

2. Residual variances are specified so that the ratio between the maximum and minimum variances is equal to 1.926 (equivalent to the standardized regression coefficient of .2). To get the residual variances, first, three numbers are drawn from the uniform distribution between 1 and 1.926. Then, the order of 1, 1.926, and the three randomly drawn three values is shuffled. This step ensures that the ratio between maximum and minimum variances is still 1.926 and these maximum and minimum variances can be the residual variances of any time points. The five shuffled values are denoted as c_1, c_2, c_3, c_4, and c_5. Next, the residual variance at the first time point is the common residual variance estimated from the data analysis. The residual variances of the second, third, fourth, and fifth time points are the residual variance at the first time point multiplied by c_2/c_1, c_3/c_1, c_4/c_1, and c_5/c_1, respectively. I draw 100 sets to search for the minimal severe misspecification.

3. The factor loadings of the linear trend are deviated such that the proportion of variance of the deviated trend explained by the linear trend is exactly equal to .90. The procedure for creating the deviated trend is explained above (setting $\rho^2 = .9$). I draw 100 sets to search for the minimal severe misspecification.

4. A quadratic trend is omitted. This quadratic trend has the mean of 0 and the variance of 0.083 (the expected value of the change when the value of quadratic trend equals one standard deviation has 90% of its variance explained by the linear trend). This quadratic factor is indepenent to the other latent variables.

Thus, the total number of misspecification sets will be 221 ($20+100+100+1$) for this growth curve model. Different global fit indices are calculated from these sets of misspecification and the set with the maximum value of each fit index are picked. Eight fit indices are used for global fit evaluation: population RMSEA, SRMR, CFI, TLI, and four correlation-based fit indices. The correlation-based fit indices detect the misspecification at the mean structure (Type B and C misspecifications). More details on the correlation-based fit indices are provided in the next section.

Regarding to the local fit evaluation, EPCs for each fixed parameters are investigated. The fixed parameters that are used for scale identification are not considered in the local fit evaluation: all measurement intercepts and the factor loadings at the first and last time points. All other fixed parameters have the ranges of trivial misspecifications as follows:

1. The range between -.2 and .2 is deemed trivial for measurement error correlations among repeated measures.

2. The ratios of the expected changed variance and the estimated variance between 0.519 (1/1.926) and 1.926 are deemed trivial.

3. The ranges of trivial misspecifications for the factor loadings at the second, third, and fourth time points are computed by the ranges of deviated expected means, $\tilde{\boldsymbol{\mu}}$. As mention above, the misspecification will be deemed trivial if the proportion of variance of $\tilde{\boldsymbol{\mu}}$ explained by $\boldsymbol{\mu}$ (ρ^2) of .90. The ranges of $\tilde{\boldsymbol{\mu}}$ can be calculated by the method described above. This step will be repeated for 100 times. The minimum and maximum values of the expected means at each time point are the range of trivial misspecification of the expected means. The minimum and maximum values of the expected means at each time point ($\tilde{\mu}_t$) are used to calculate the minimum and maximum values of the EPCs for the factor loadings as follows:

$$\tilde{\mu}_t = \tilde{\lambda}_{t1}\hat{\alpha}_1 + \tilde{\lambda}_{t2}\hat{\alpha}_2 \tag{4.7}$$

4.2.3 Additional Fit Indices for Growth Curve Models

The fit indices used in the CFA model in Study 1 detect the misspecification at the covariance matrix. In growth curve model, however, the misspecification at the mean structure needs to be detected by fit indices as well. Wu & West (2013) showed that four correlation-based fit indices can be used to detect the misspecification at the marginal means and the misspecification in predicting individual responses (or misspecification at the conditional means). Marginal Pseudo R^2 represents the squared correlation between estimated marginal means and the observed individual responses:

$$Marginal\,Pseudo\,R^2 = \left(r_{Y\hat{\mu}}\right)^2 = \frac{\left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)'\left(\hat{\boldsymbol{\mu}}_i - \hat{y}\mathbf{1}_i\right)\right]}{\left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)'\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)\right]\cdot\left[\sum_{i=1}^{n}\left(\hat{\boldsymbol{\mu}}_i - \hat{y}\mathbf{1}_i\right)'\left(\hat{\boldsymbol{\mu}}_i - \hat{y}\mathbf{1}_i\right)\right]}. \tag{4.8}$$

where Y represents the observed individual responses and $\hat{\mu}$ represents the predicted marginal means. n is the number of participants. $\bar{y}$ represents the grand mean of the observed responses across participants and across time points. $\hat{y}$ represents the grand mean of the estimate responses ($\hat{\mathbf{y}}_i$) across participants and across time points. Conditional Pseudo R^2 represents the squared correlation between estimated conditional means and the observed individual responses.

$$Conditional\,Pseudo\,R^2 = \left(r_{Y\hat{Y}}\right)^2 = \frac{\left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)'\left(\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_i\right)\right]}{\left[\sum_{i=1}^{n}\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)'\left(\mathbf{y}_i - \bar{y}\mathbf{1}_i\right)\right]\cdot\left[\sum_{i=1}^{n}\left(\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_i\right)'\left(\hat{\mathbf{y}}_i - \hat{y}\mathbf{1}_i\right)\right]}. \tag{4.9}$$

where $\hat{Y}$ represents the predicted conditional means.

In a perfect correlation, two variables measuring the same thing would produce a scatterplot that individual points follow a 45-degree straight line. Concordance Correlation (CCC) represents the degree of deviations of individual points from a 45-degree straight line. Conditional CCC

assesses the degree to which the predicted individual responses reproduce the observed individual responses:

$$Conditional\,CCC = 1 - \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)'(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)}{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{y}\mathbf{1}_i)'(\boldsymbol{y}_i - \bar{y}\mathbf{1}_i) + \sum_{i=1}^{n}(\hat{\boldsymbol{y}}_i - \hat{y}\mathbf{1}_i)'(\hat{\boldsymbol{y}}_i - \hat{y}\mathbf{1}_i) + N(\bar{y} - \hat{y})^2}. \quad (4.10)$$

where n is the number of individuals and N is the total number of observations ($n \times 5$ in this growth curve model). Average CCC assesses the extent to which the predicted marginal means reproduce the observed individual responses:

$$Average\,CCC = 1 - \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)'(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)}{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{y}\mathbf{1}_i)'(\boldsymbol{y}_i - \bar{y}\mathbf{1}_i) + \sum_{i=1}^{n}(\hat{\boldsymbol{\mu}}_i - \hat{y}\mathbf{1}_i)'(\hat{\boldsymbol{\mu}}_i - \hat{y}\mathbf{1}_i) + N(\bar{y} - \hat{y})^2}. \quad (4.11)$$

Type B and C misspecifications represent the misspecification on the mean structure. Type C misspecification would alter the marginal means so all four fit indices could detect this misspecification. Type B misspecification, however, does not alter the marginal means because the mean of the quadratic factor is 0. Thus, the marginal pseudo R^2 and average CCC cannot detect this misspecification. Conditional pseudo R^2 and conditional CCC should be able to detect this misspecification because they detect the discrepancy between the observed and the predicted individual responses. The global fit evaluation from the unified approach provides the expected range of these four correlation-based fit indices under trivial or severe misspecifications.

4.2.4 Simulation Analysis

Because the unified approach is the three-outcome method, two outcome variables are used: the proportion of inconclusive (or underpowered) results and the proportion of model rejection among conclusive results. Similar to Study 1, the rejection rate is coded as missing if the proportion of inconclusive results is greater than .90.

Similar to Study 1, the tables of η^2s for rejection rate and the proportion of inconclusive results

are provided. Four design factors include the size of error variances, sample size, the degree of misspecifications, and the type of misspecifications. Similar to Study 1, I did not include the results from Level 0 degree of misspecification. The factors with an η^2 of .03 or higher are deemed non-negligible factors. The patterns of the non-negligible effects are described in texts or by tables. If the desired effects are related to the degree of misspecification, the results with Level 0 degree of misspecification are included. Given that the main effect of the degree of misspecification is the primary research question of the study. The tables of the averages of the rejection rates and the proportion of inconclusive results across the degree of misspecification are provided even if the effects were negligible. If the effect of the degree of misspecification depends on other factors (i.e., a two-way or higher interaction), the table describing the higher order interaction is shown instead. The tables mentioned above are used to answer the research questions as follows.

4.2.4.1 Rejection Rates when the Unified Approach Provides Conclusive Results

Appropriate Rejection Rates for Varying Degrees of Misspecification. The rejection rate for trivial misspecification should be 0 and the rejection rate for severe misspecification should be 1. Three results are used to indicate whether this desired characteristic is satisfied. First, the degree of model misspecification should influence the rejection rate so the η^2 of the main effect of the degree of model misspecification should be greater than .03. Second, the rejection rates of trivial misspecification conditions should be low. Third, the rejection rates of severe misspecification conditions should be high.

Rejection Rates Are Not Influenced by Types of Misspecification, Model Characteristic, and Sample Size The unified approach should have low η^2s ($<$ .03) on the main and interaction effects involving the type of misspecification, the amount of error variance, or sample size.

4.2.4.2 The Properties of the Unified Approach

The Pattern of the Proportions of Inconclusive Results. The proportion of inconclusive results should be higher when (a) sample size is low and (b) when the degree of misspecification is close

to the level of maximal trivial misspecification (i.e., the degree of model misspecification is Level 1 or 2) holding sample size constant.

The Congruency between Global and Local Model Evaluation. I use a contingency table to examine the interaction between the results from both methods. The contingency table will be used to check whether global or local fit evaluation is unnecessary to be used. Furthermore, I investigate the situation where one method indicates trivial misspecification while the other indicates severe misspecification. This situation should not occur.

Chapter 5

Results

5.1 Study 1

The first simulation study compares the performance of all model evaluation methods and evaluates the characteristics of the unified approach. The model evaluation methods examined and the abbreviations used in the tables in this chapter are as follows:

1. A one-size-fit-all cutoff for RMSEA (RMSEA)
2. A one-size-fit-all cutoff for CFI (CFI)
3. A one-size-fit-all cutoff for TLI (TLI)
4. A one-size-fit-all cutoff for SRMR (SRMR)
5. The combination of all one-size-fit-all fit indices cutoffs. A model is rejected if any of (1) - (4) indicates model rejection (OVCUT)
6. Test of close fit and not close fit (CLOSE)
7. The modification indices and power approach (MIPOW)
8. The PPP cutoff for the Bayesian analysis with informative priors on cross loadings (LOAD, PPP)

9. The combination of the PPP cutoff and the zero coverage of the credible intervals from cross loadings informative priors (LOAD, ZERO)

10. The PPP cutoff for the Bayesian analysis with informative priors on error covariances (ERR, PPP)

11. The combination of the PPP cutoff and the zero coverage of the credible intervals from error covariances informative priors (ERR, ZERO)

12. The simulation approach combining all three types of trivial misspecification. A model is rejected if any of the three types of misspecification provide model rejection (SIM)

13. The unified approach (UNIFIED)

I do not report the results from different types of maximal trivial misspecification in the simulation approach because they follow the same pattern. The only result from the simulation approach reported here is the combination of the results from all types of maximal trivial misspecification. I do not combine the results from Bayesian analysis because different priors and different model evaluation methods led to different results.

Table 5.1: The η^2s of the Effects of the Design Conditions on the Rejection Rates for Study 1

Factors	RMSEA	CFI	TLI	SRMR	OVCUT	CLOSE	MIPOW	Bayesian Analysis				SIM	UNIFIED
								LOAD, PPP	LOAD, ZERO	ERR, PPP	ERR, ZERO		
TYPEMIS	.018	.024	**.038**	.025	.004	**.045**	.003	**.296**	**.124**	.027	**.154**	**.046**	.001
N	.001	.003	.001	.007	.002	.000	**.053**	.006	**.053**	**.032**	.004	.025	.011
LOAD	.001	**.035**	.020	.024	.000	.009	.003	.000	.000	**.080**	**.034**	.000	.003
SEVERE	**.731**	**.728**	**.720**	**.689**	**.739**	**.695**	**.236**	**.219**	**.423**	**.114**	**.060**	**.624**	**.612**
LEVELMIS	.000	.000	.000	.000	.000	.000	**.206**	**.047**	.026	**.248**	**.117**	**.039**	**.100**
ITEMS	**.045**	.004	.007	.000	.004	.013	.003	.002	.006	**.089**	**.083**	.001	.004
TYPEMIS : N	.000	.000	.000	.000	.000	.000	.000	**.037**	.010	.002	.026	.001	.000
TYPEMIS : LOAD	.000	.006	.000	.022	.019	.003	.000	.000	.001	.001	.014	.001	.000
TYPEMIS : SEVERE	.000	.000	.000	.000	.000	.000	.000	.013	.000	.013	.001	.004	.005
TYPEMIS : LEVELMIS	.000	.000	.000	.000	.000	.000	.000	.013	.020	.000	**.057**	.001	.000
TYPEMIS : ITEMS	.009	.000	.000	.001	.000	.004	.002	.002	.001	.003	.006	.001	.002
N : LOAD	.000	.000	.001	.001	.001	.000	.001	.000	.001	.012	.000	.000	.003
N : SEVERE	.000	.002	.003	.000	.005	.000	**.047**	.025	.025	.000	.002	.008	.000
N : LEVELMIS	.000	.000	.000	.000	.000	.000	.006	.002	.015	.000	.004	.009	.005
N : ITEMS	.000	.000	.000	.001	.001	.000	.003	.000	.001	.000	.008	.000	.000
LOAD : SEVERE	.000	.002	.003	.000	.002	.000	.000	.000	.001	**.032**	.001	.000	.000
LOAD : LEVELMIS	.000	.000	.000	.000	.000	.000	.001	.000	.003	.027	.000	.000	.007
LOAD : ITEMS	.000	.001	.000	.000	.000	.000	.000	.000	.000	.003	.006	.000	.000
SEVERE : LEVELMIS	.000	.000	.000	.000	.000	.000	**.132**	**.068**	.000	.014	.005	.015	**.128**
SEVERE : ITEMS	.000	.000	.000	.000	.000	.000	.004	.001	.005	.002	**.073**	.000	.000
LEVELMIS : ITEMS	.000	.000	.000	.000	.000	.000	.001	.000	.000	.001	.015	.000	.000

Note.

1. The boldface numbers represent the η^2s $\geq .03$.
2. All interactions higher than two ways are not presented here because their η^2s $< .03$.
3. The abbrevations for the model evaluation methods are provided at the beginning of this chapter.
4. TYPEMIS = Type of misspecifications, N = Sample size, LOAD = The size of target factor loadings, SEVERE = The degree of misspecification, LEVELMIS = The level of trivial misspecification, ITEMS = The number of items

Table 5.2: The Rejection Rates and the Proportions of Inconclusive Results for Each Model Evaluation Method Classified by the Level of Maximal Trivial Misspecification and the Degree of Misspecification for Study 1

Level of Trivial Misspecification	Level 1				Level 2			
Degree of Model Misspecification	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3
Classification	Trivial	Cutoff	Severe	Severe	Trivial	Trivial	Cutoff	Severe
				Rejection Rates				
RMSEA	.011	.018	.371	.999	.011	.018	.371	.999
CFI	.037	.055	.533	1.000	.037	.055	.533	1.000
TLI	.049	.080	.672	1.000	.049	.080	.672	1.000
SRMR	.016	.032	.420	.981	.016	.032	.420	.981
OVCUT	.058	.098	.844	1.000	.058	.098	.844	1.000
CLOSE	.003	.009	.662	1.000	.003	.009	.662	1.000
MIPOW	.751 (.157)	.892 (.073)	1.000 (.004)	1.000 (.000)	.253 (.137)	.253 (.130)	.662 (.046)	1.000 (.000)
Bayesian, LOAD, PPP	.004	.109	.302	.904	.005	.112	.282	.339
Bayesian, LOAD, ZERO	.031	.307	.730	.985	.020	.151	.609	.846
Bayesian, ERR, PPP	.384	.402	.599	.972	.079	.080	.098	.343
Bayesian, ERR, ZERO	.681	.846	.956	1.000	.354	.556	.779	.803
SIM	.017	.273	.833	1.000	.002	.005	.572	.999
UNIFIED	.000 (.840)	.740 (.953)	1.000 (.462)	1.000 (.001)	.000 (.343)	.000 (.356)	.661 (.874)	1.000 (.102)

Note. The abbrevations for the model evaluation methods are provided at the beginning of this chapter. The numbers in the parentheses represent the proportion of inconclusive results.

5.1.1 Convergence Rates

All model evaluation methods except Bayesian analysis are based on maximum likelihood estimation. Almost all conditions using maximum likelihood had convergence rate greater than 98%, except that for the conditions with sample size of 125, the size of target loadings of .5, 8 items, and Level 3 misspecification in factor loadings, the convergence rate was 88%.

The convergence rates of Bayesian analysis were lower than ones in maximum likelihood. The convergence rates were lower when the degree of misspecification increased, especially for Level 3 misspecification. Within the conditions for Level 3 misspecification, the Bayesian analysis with informative priors on cross loadings had low convergence rates when the misspecification was in measurement error correlations and the size of the target factor loadings was .5 (12%). On the contrary, the Bayesian analysis with informative priors on error covariances had low convergence rates when the misspecification was in factor correlation and the size of the target factor loadings was .7 (29%).

5.1.2 The Comparison between Model Evaluation Methods

5.1.2.1 Rejection Rate for Model Misspecification and Level of Trivial Misspecification

This section investigates the performance of all model evaluation methods in detecting trivial and severe misspecifications. The ANOVA was used to investigate the η^2s of the interaction effect between the degree of misspecification (in data-generating models) and the level of trivial misspecification (used in the unified approach). As shown in Table 5.1, three methods had nonnegligible η^2s of the interaction: the modification indices and power approach, the PPP method in Bayesian analysis with informative priors on cross loadings, and the unified approach. All of the other methods had rejection rates sensitive to the main effect of the degree of model misspecification or the level of trivial misspecification but they were not sensitive to the interaction effect.

Table 5.2 provides the rejection rates classified by the degree of model misspecification and the level of trivial misspecification. The rejection rates for the trivial misspecification conditions are expected to be close to 0. As mentioned in Chapter 4, model evaluation methods with rejection rates less than .1 are to be labelled as correctly retaining models. The Bayesian approach with informative priors on error covariances (both ERR, PPP and ERR, ZERO) and the modification indices and power approach had the rejection rates over .1, especially when the level of maximal trivial misspecification was 0.1. Other methods had rejection rates less than .1 for all trivial misspecification conditions.

Ideal rejection rates for the severe misspecification conditions are close to 1. The model evaluation methods with rejection rates of .9 or higher are deemed correctly rejected models. Only the unified approach provided the rejection rate of 1 in all severe misspecification conditions. Other methods had a rejection rate lower than .9 for at least one severe misspecification condition. Finally, the rejection rates for the cutoff conditions are expected to range between .1 and .9. All methods provided the desired results.

5.1.2.2 The Effect of Types of Misspecification

As shown in Table 5.1, All methods except the TLI cutoff, the test of close fit and not close fit, the PPP method from the Bayesian analysis with informative priors on cross loadings (LOAD, PPP), and the PPP method with the zero coverage of the credible intervals from informative priors on error covariances (ERR, ZERO) had negligible η^2s for all main and interaction effects involving types of misspecification. The TLI cutoff had lower rejection rates when the misspecification was in factor correlations (.44) but higher rejection rates when the misspecification was in cross loadings or error correlations (.65). The test of close fit and not close fit approach provided the rejection rates of .41, .59, and .66 for the misspecifications in factor correlation, cross loadings, and error correlations, respectively.

The PPP method with cross loadings priors had nonnegligible η^2 for the interaction effect between types of misspecification and sample size. If the misspecification was in factor correlations, the rejection rates slightly decreased when sample size increased (.14 and .02 in sample size of 125 and 4000, respectively). If the misspecification was in cross loadings, the rejection rates were in between .15 and .19 in all sample sizes. However, If the misspecification was in error correlations, the rejection rates from the PPP approach increased when sample size increased. The rejection rates were .44, .60, .68, .74, .89, and .99 in sample sizes of 125, 250, 500, 1000, 2000, and 4000, respectively.

The PPP method with the zero coverage of the credible intervals from error covariances informative priors had nonnegligible interaction effect of the type of misspecification and the level of trivial misspecification. In general, it had lower rejection rates when the misspecification was in factor correlations (.40 - .86) but higher rejection rates when the misspecification was in cross loadings or error correlations (.85 - .97). The level of trivial misspecification influenced the rejection rates of the model with the misspecification in factor correlation. The rejection rate was .40 in the trivial misspecification at Level 1 but .86 in the trivial misspecification at Level 2.

5.1.2.3 The Effect of Model Characteristics

As shown in Table 5.1, the number of items had no effect on any of the evaluation methods except the RMSEA cutoff approach and the Bayesian approach using informative priors on error covariances (both ERR, PPP and ERR, ZERO). For the RMSEA cutoff approach, the rejection rate was lower when the number of items increased (rejection rates were .56 and .36 for 8 and 16 items, respectively). In contrast, for the Bayesian analysis with error correlations, the rejection rates were higher when the number of items increased. The rejection rates for the PPP method were .27 and .55 for 8 and 16 items repectively. When the PPP method and the zero coverage of nontarget credible intervals are combined, the effect of the number of items on rejection rates depended on the degree of misspecification. When the degree of misspecification was low (Level 1), the rejection rates were .49 and .91 for 8 and 16 items, respectively. However, when the degree of misspecification was high (Level 3), the rejection rates were .90 and .90 for 8 and 16 items, respectively.

As shown in Table 5.1, the size of factor loadings had a negligible effect on all model evaluation methods except the CFI cutoff approach and all model evaluation methods using the Bayesian analysis with informative priors on error covariances (both ERR, PPP and ERR, ZERO). For the CFI cutoff, the rejection rate was lower when the size of loadings increased. The rejection rates were .61 and .45 for the loadings of .5 and .7, respectively. For the Bayesian analysis with error covariances informative priors, the effect of the size of loadings was influenced by the degree of misspecification. When the degree of misspecification was low (Level 1), the rejection rates were .46 and .03 for the size of factor loadings of .5 and .7, respectively. However, when the degree of misspecification was high (Level 3), the rejection rates were .65 and .64 for the size of factor loadings of .5 and .7, respectively. The rejection rates using the combination of the PPP method and the zero coverage were .88 and .76, respectively.

5.1.2.4 The Effect of Sample Size

Only the modification indices and power approach and all model evaluation methods using Bayesian analysis had non-negligible η^2s for either main or interaction effects involving sample size. Sample size and the degree of misspecification interacted with each other to affect the performance of the modification indices and power approach. As shown in Table 5.3, the rejection rates were close to 1.0 regardless of sample size for Level 3 degree of misspecification. However, for Level 1 degree of misspecification, the rejection rates were close to 1 for small sample sizes and close to 0 for larger sample sizes. The influence of sample size was inconsistent across the methods using Bayesian analysis. Increase in sample size increased the rejection rates for the methods using cross loadings informative priors whereas reduced rejection rates for the methods using error covariances informative priors.

In sum, the unified approach satisfied all of the desired properties. As shown in Table 5.1, the effect size of the interaction between the degree of misspecification and the level of trivial misspecification was non-neglibible for the unified approach. As shown in Table 5.2, when the level of trivial misspecification was 0.1, the rejection rates were 0 for the degree of misspecification lower than 0.1 and the rejection rates were 1 for the degree of misspecification higher than 0.1. When the level of trivial misspecification was 0.3, the rejection rates were 0 for the degree of misspecification lower than 0.3 and the rejection rates were 1 for the degree of misspecification higher than 0.3. Sample size, type of misspecification, number of items, and size of factor loadings did not influence the rejection rates from the unified approach. Because of these properties, the unified approach outperformed the other model evaluation methods. However, the unified approach tended to provide inconclusive results in some conditions (e.g., under small sample size). The

Table 5.3: The Rejection Rates of the Modification Indices and Power Approach Classified by the Degree of Misspecification and Sample Size for Study 1

Degree of Model Misspecification	Sample Size	Rejection Rate
0	125	1.000
	250	.752
	500	.508
	1000	.500
	2000	.253
	4000	.001
1	125	1.000
	250	.753
	500	.507
	1000	.500
	2000	.403
	4000	.274
2	125	1.000
	250	.900
	500	.795
	1000	.768
	2000	.762
	4000	.760
3	125	1.000
	250	1.000
	500	1.000
	1000	1.000
	2000	1.000
	4000	1.000

performance of the unified approach is examined in more detailds in the next section.[1]

5.1.3 The Properties of the Unified Approach

5.1.3.1 The Pattern of the Proportions of Inconclusive Results

The η^2s investigating the influence of all design conditions on the proportion of inconclusive results are shown in Table 5.4. The proportion of inconclusive results was influenced by sample size and the interaction between the degree of misspecification and the level of trivial misspecification. Table 5.5 provides the more detailed version of Table 5.4 that the results of the unified approach are further classified by sample size. The proportion of inconclusive results was lower when sample size increased. The bottom part of Table 5.2 shows the proportions of inconclusive results classified by the degree of misspecification and the level of trivial misspecification. The proportion of inconclusive results was the highest for the cutoff conditions. In addition, the proportion of inconclusive results for severe misspecification was lower than that for trivial misspecification.

5.1.3.2 Congruency between Global and Local Model Evaluation

Table 5.6 shows the contingency table between the results from the global and local fit evaluations. Each cell represents the proportion of each outcome from all replications. The global fit evaluation yielded 81% of inconclusive results and 19% of severe misspecification, as well as less than 1% of trivial misspecification. The local fit evaluation yielded 15% of trivial misspecification, 30% of severe misspecification, as well as 55% of inconclusive results (including underpowered). These two methods complemented each other. Some inconclusive results from global fit evaluation approach

[1]The unified approach calculated rejection rates differently from the other model evaluation methods because the rejection rates were calculated from conclusive replications only. Appendix E provides the results of the other model evaluation methods when their rejection rates were calculated based on the conclusive replications from the unified approach. The results from the appendix indicated that the modification indices and power approach and the simulation approach had desirable properties similar to the unified approach. Thus, if the unified approach provided conclusive results, the results from the modication indices and power approach were reliable. However, the results from both approaches should be interpreted carefully if the unified approach provided inconclusive results.

Table 5.4: The η^2s of the Effects of the Design Conditions on the Proportions of Inconclusive Results for Study 1

Factors	MIPOW	UNIFIED
TYPEMIS	.002	.000
N	**.054**	**.085**
LOAD	.000	.002
SEVERE	**.120**	**.284**
LEVELMIS	.019	.001
ITEMS	**.078**	.014
TYPEMIS : N	.001	.000
TYPEMIS : LOAD	.000	.000
TYPEMIS : SEVERE	.001	.000
TYPEMIS : LEVELMIS	.000	.000
TYPEMIS : ITEMS	.001	.000
N : LOAD	.000	.000
N : SEVERE	**.047**	.011
N : LEVELMIS	.025	.001
N : ITEMS	.028	.003
LOAD : SEVERE	.000	.000
LOAD : LEVELMIS	.001	.000
LOAD : ITEMS	.001	.000
SEVERE : LEVELMIS	.009	**.095**
SEVERE : ITEMS	**.076**	.000
LEVELMIS : ITEMS	.005	.000

Note. The boldface numbers represent the η^2s $\geq$ .03. All interactions higher than two ways are not presented here because their η^2s $<$ .03. MIPOW = The modification indices and power approach, UNIFIED = The unified approach, TYPEMIS = Type of misspecifications, N = Sample size, LOAD = The size of target factor loadings, SEVERE = The degree of misspecification, LEVELMIS = The level of trivial misspecification, ITEMS = The number of items.

Table 5.5: The Rejection Rates and The Proportions of Inconclusive Results Classified by Sample Size, Degree of Model Misspecification, and Level of Trivial Misspecification for the unified approach in Study 1

Level of Trivial Misspecification	Level 1				Level 2				Marginal Average
Degree of Model Misspecification	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3	
Classification	Trivial	Cutoff	Severe	Severe	Trivial	Trivial	Cutoff	Severe	
Rejection Rates									
Sample Size									
125	NA	NA	NA	1.000	NA	NA	NA	1.000	1.000
250	NA	NA	NA	1.000	NA	NA	NA	1.000	1.000
500	NA	NA	1.000	1.000	.000	.000	.662	1.000	.610
1000	NA	NA	1.000	1.000	.000	.000	.589	1.000	.598
2000	NA	NA	1.000	1.000	.000	.000	.567	1.000	.595
4000	.000	.661	1.000	1.000	.000	.000	.584	1.000	.531
Marginal Average	.000	.661	1.000	1.000	.000	.000	.601	1.000	
The Proportions of Inconclusive Results									
125	1.000	.999	.942	.005	.999	1.000	.971	.496	.801
250	1.000	1.000	.930	.000	.879	.912	.950	.115	.723
500	.988	.952	.540	.000	.174	.214	.894	.003	.471
1000	.999	.950	.354	.000	.004	.009	.853	.000	.396
2000	.947	.933	.008	.000	.000	.000	.811	.000	.337
4000	.104	.885	.000	.000	.000	.000	.766	.000	.219
Marginal Average	.840	.953	.462	.001	.343	.356	.874	.102	

Note. The abbrevations for the model evaluation methods are provided at the beginning of this chapter. NA = The proportion of inconclusive results were greater than .90.

Table 5.6: The Contingency Table of the Proportions of the Results from Global and Local Fit Evaluation in the Unified Approach

Global \Local	Trivial	Underpowered	Inconclusive	Severe	Marginal Proportion
Trivial	**.003**	.000	.000	.000	*.004*
Inconclusive	.145	**.220**	**.251**	.188	*.805*
Severe	.000	.082	.000	**.109**	*.191*
Marginal Proportion	*.149*	*.302*	*.252*	*.297*	

Note. The boldface numbers represent the matched results between both global and local fit evaluations. The italicized numbers represent the marginal proportions of the results from each method.

were classified as trivial or severe misspecification by local fit evaluation. Some underpowered results from local fit evaluation were classified as severe misspecification in the global fit evaluation. Across all replications, the proportions of inconclusive results from local or global fit evaluation alone (55% or 81%, respectively) were higher than the proportions of inconclusive results from the two approaches combined (47%).

The proportion of inconclusive results from local fit evaluation was highly dependent on the level of maximal trivial misspecification and sample size. If the level of maximal trivial misspecification was low, the range of trivial misspecification would be narrow. The confidence interval of EPC had to be narrow as well to get a conclusive result so a large sample size was needed. In this case, global fit evaluation could complement the local fit evaluation. Lower level of maximal trivial misspecification increased the chance of rejecting a model in global fit evaluation rather than providing the higher chance of inconclusive results in local fit evaluation.

On the other hand, to reject a model, global fit evaluation required the model to have a very severe misspecification. If the degree of misspecification was not extremely severe, local fit evaluation had a higher power than the global fit evaluation in detecting the misspecification, especially with large sample sizes. Local fit evaluation was also able to correctly classify a trivially misspecified model whereas global fit evaluation had extremely low power to detect trivial misspecification.

5.2 Study 2

The second study evaluates the performance of the unified approach for growth curve models. Similar to Study 1, I examine whether the unified approach can correctly classify trivially or severely misspecified models. The properties of the unified approach are also investigated in this study.

5.2.1 Convergence Rates

Most conditions had convergence rates higher than 99%. However, the convergence rates were highly dependent on the degree of misspecification and the type of misspecification, as shown in

Table 5.7: The average convergence rates classified by the degree of misspecification and the type of misspecification.

Degree of Misspecification	Type of Misspecification	Convergence Rate
Level 0		1.00
Level 1	Type A: Omitted Autoregressive Regression	1.00
	Type B: Omitted Quadratic Factor	1.00
	Type C: Nonlinear Change	1.00
	Type D: Unequal Error Variances	1.00
Level 2	Type A: Omitted Autoregressive Regression	1.00
	Type B: Omitted Quadratic Factor	1.00
	Type C: Nonlinear Change	1.000
	Type D: Unequal Error Variances	.96
Level 3	Type A: Omitted Autoregressive Regression	1.00
	Type B: Omitted Quadratic Factor	1.00
	Type C: Nonlinear Change	.09
	Type D: Unequal Error Variances	.03

Table 5.7. Specifically, when the degree of misspecification was Level 3 and the misspecification was in nonlinear change (type C misspecification) or unequal error variances (type D misspecification), the convergence rates were lower than 10%.

5.2.2 Rejection Rates when the Unified Approach Provides Conclusive Results

When the unified approach yielded conclusive results, the accuracy of model rejection or retention is investigated. The first way to investigate the accuracy is to examine the η^2 associated with the main effect of the degree of misspecification. As shown in Table 5.8, the interaction effect between the degree of misspecification and sample size was substantial. Table 5.9 provides the average rejection rates across these conditions. When sample size was small and the degree of misspecification was low, the unified method provided high proportions of inconclusive results so these cells were empty. If sample size was higher (> 500) and the degree of misspecification was low, the rejection rates were close to 0. However, when sample size was 500 and the degree

Table 5.8: The η^2s of the Effects of the Design Factors on the Rejection Rates and the Proportion of Inconclusive Results for Study 2

Factors	Rejection Rates	Proportions of Inconclusive Results
TYPEMIS	.010	.009
N	**.149**	**.237**
ERRVAR	.028	**.032**
SEVERE	**.347**	**.284**
TYPEMIS : N	.000	.002
TYPEMIS : ERRVAR	.004	.001
TYPEMIS : SEVERE	.011	.011
N : ERRVAR	.004	.000
N : SEVERE	**.067**	.008
ERRVAR : SEVERE	.023	.011

Note. The boldface numbers represent the η^2s $\geq$ .03. All interactions higher than two ways are not shown here because their η^2s < .03. TYPEMIS = Type of misspecifications, N = Sample size, ERRVAR = The amount of error variances, SEVERE = The degree of misspecification.

of misspecification was Level 1, the rejection rates were .33, which was higher than the desired rejection rate (< .10). For severe misspecification conditions, the rejection rates were close to 1. However, the rejection rates tended to decrease in a small amount when sample size increased.

As shown in Table 5.8, the amount of error variances and type of misspecification did not have a significant effect on the rejection rates from the unified approach. Increase in sample size, however, slightly decreased rejection rates as described above.

5.2.3 The Properties of the Unified Approach

5.2.3.1 The Pattern of the Proportions of Inconclusive Results

The η^2s associated with all design factors on the proportion of inconclusive results are shown in Table 5.8. The proportion of inconclusive results was influenced by sample size, the amount of error variances, and the degree of misspecification. As expected, when sample size increased, the proportion of inconclusive results decreased (.85, .63, .48, and .28 for the sample sizes of 125, 250, 500, and 1000, respectively). When the amount of error variances increased, the proportion

Table 5.9: The Rejection Rates of the Unified Approach Classified by the Degree of Misspecification and Sample Size for Study 2

Degree of Misspecification	Sample Size	Rejection Rates	Proportions of Inconclusive Results
0	125	NA	1.000
	250	NA	.998
	500	.001	.689
	1000	.000	.091
1	125	NA	1.000
	250	NA	.978
	500	.333	.770
	1000	.091	.391
2	125	NA	.908
	250	1.000	.660
	500	.909	.487
	1000	.834	.305
3	125	1.000	.599
	250	1.000	.181
	500	1.000	.108
	1000	.998	.088

Note. NA = The proportion of inconclusive results were greater than .90.

of inconclusive results increased (.46, .60, and .63 for the error variances of 0.5, 2, and 3.5, respectively). Finally, across the degree of misspecification, the order of the proportion of inconclusive results from the highest to the lowest values was Levels 1 (.78), 0 (.69), 2 (.59), and 3 (.25). Therefore, the proportion of inconclusive results was the highest when the degree of misspecification was close to the level of maximal misspecification.

5.2.3.2 Congruency between Global and Local Model Evaluation

Table 5.10 shows the contingency table between the results from the global and local fit evaluations. Each cell represents the proportion of each outcome among all replications. Global fit evaluation yielded 95% of inconclusive results and 5% of severe misspecification results. The local fit evaluation yielded 10% of trivial misspecification, 26% of severe misspecification, as well as 64% of all inconclusive results (including underpowered). Similar to Study 1, these two methods complemented each other. Across all replications, the proportions of inconclusive results from local and global fit evaluation alone (95% and 64%, respectively) were higher than the proportions of inconclusive results from the two methods combined (62%).

Table 5.10: The Contingency Table of the Results from Global and Local Fit Evaluation in the Unified Approach

Global \Local	Trivial	Underpowered	Inconclusive	Severe	Marginal Proportion
Inconclusive	.100	**.294**	**.335**	.224	*.953*
Severe	.000	.013	.000	**.033**	*.047*
Marginal Proportion	*.100*	*.307*	*.335*	*.257*	

Note. The boldface numbers represent the matched results between both global and local fit evaluations. The italicized numbers represent the marginal proportions of the results from each method.

The proportion of inconclusive results in local fit evaluation was highly dependent on the level of trivial misspecification, sample size, and type of misspecification. For highly severe misspecification (Level 3), if the sample size was low (125) and the misspecifications were the deviated trend or different error variances, global fit evaluation indicated severe misspecification but local fit evaluation yielded inconclusive results. On the other hand, if sample size was large and the misspecifications were the misspecified autoregressive regression or unspecified quadratic factors, the local fit evaluation indicated severe misspecification but the global fit evaluation provided inconclusive results. If the degree of misspecification was lower and sample size was large (>= 250), global fit evaluation yielded inconclusive results more often than the local fit evaluation. The gap was wider if sample size increased.

Chapter 6

Discussion and Conclusion

The goal of this dissertation is to develop a unified approach for model evaluation in structural equation modeling. In the result chapter, I found that the unified approach outperformed other model evaluation methods in evaluating model fit in confirmatory factor analysis. The unified approach performed well in evaluating model fit in growth curve models also. In this chapter, I summarize the desirable properties of the unified approach. I further discuss how the problems of the current practices of model evaluation discussed in Chapter 1 are solved by the unified approach. Next, the limitations of the unified approach are discussed, such as large sample size requirement, subjectivity, and the possibility of inconclusive results. The suggestions for dealing with the limitations are provided. Then, I discuss the possible extensions of the unified approach, including sample size estimation (or power analysis) and nested model comparison. Finally, the limitations of the simulation studies are discussed.

6.1 The Performance of the Unified Approach

From the simulation studies in the previous chapter, the unified approach appropriately rejected severely misspecified models and retained trivially misspecified models when it provided conclusive results. The accuracy of the unified approach was better than the other model evaluation methods. That is, the rejection rates for severely misspecified models were close to 1 and those for

trivially misspecified models were close to 0. In addition, the rejection rates were appropriately adjusted for the level of maximal trivial misspecification. Furthermore, its performance was not influenced by model characteristics or sample size.

There were three main reasons why the unified approach provided appropriate rejection rates. First, when there is not enough information, the unified approach leads to inconclusive results. According to the simulation studies, it tends to provide inconclusive results in three situations: (a) low sample size, (b) when the degree of misspecification is close to the level of maximal trivial misspecification, and (c) high error variances (in growth curve models). Low sample size increases the width of the confidence intervals in local fit evaluation. Hence, the width of the confidence intervals could be larger than the range of trivial misspecification, leading to underpowered results. Thus, researchers should ensure that the resulting confidence intervals should be narrow enough (e.g., by collecting larger sample) so that the underpowered results are unlikely to occur. When the degree of misspecification is close to the level of the maximal trivial misspecification, the chance that the resulting confidence intervals and the ranges of trivial misspecification are overlapped is high, leading to inconclusive results. Finally, larger error variances led to more inconclusive results. A possible explanation is that, when error variances are larger, the means of measurement errors (i.e., the standard errors of the measurement error means) have a larger variance as well. Thus, the standard errors of the means of each time point are higher and the uncertainty of the factor loadings for the linear slope is larger (see Equation 4.3). Therefore, the confidence intervals of the EPCs of factor loadings are wider leading to higher chance of underpowered results.

The second reason is that the unified approach takes all model characteristics and sample size into account. When data are generated in the global fit evaluation, the target model and their parameter values are used to generate data with the same sample size as the original data. In the local fit evaluation, the statistics from model estimation are used so all model characteristics and sample size are also taken into account. Because of this fact, the unified approach is less sensitive to model characteristics and sample size. Although sample size influences the rejection rates in the second simulation study, the amount of effect is negligible.

The third reason is that the unified approach allows users to define their own maximal trivial misspecification. Different types of trivial misspecification can be also specified simultaneously. The defined maximal trival misspecifications are added to the data-generating models in the global fit evaluation. Then, the expected ranges of fit indices from the model with trivial or severe misspecification are calculated. Thus, the cutoffs for fit indices are derived based on researchers' definition of maximal trivial misspecification. The global fit evaluation provides two cutoffs–one for retaining and one for rejecting a model–as well as the range of the values for each fit index for which the model could be trivially or severely misspecified, which contradicts to the traditional use of one-size-fit-all cutoffs. Local fit evaluation also allows users to define maximal trivial misspecification for each fixed parameter. This feature allows researchers to detect all types of misspecification according to the researchers' definitions of maximal trivial misspecification.

6.2 Does the Unified Approach Fix the Problems of the Current Practices of Model Evaluation?

In the first chapter, I discussed three problems of the current practices in model evaluation, which are mainly related to the use of one-size-fit-all cutoffs. In this section, I discuss whether the unified approach can solve the problems. The first problem is that the performance of fit indices cutoffs depends on model characteristics, sample size, model type, and data distribution. According to the simulation studies, the unified approach is not subject to the problem. When the unified approach provides conclusive results, its performance is good, regardless of model characteristics or sample size (except negligibly influenced by sample size in growth curve modeling). Sample size, however, may still influence the possibility that the unified approach generates inconclusive results. The second problem is that the derivation of the one-size-fit-all cutoffs is arbitrary. The unified approach does not provide a fixed cutoff for a fit index. Rather, it provides a method to tailor the cutoff to user-defined maximal trivial misspecification and alpha level. It is possible that users' choices on the alpha level or the level of maximal trivial misspecification is wrong. I will

discuss this issue further in the next section. The third problem is that many fit indices may lead to inconsistent results. To solve the problem, the unified approach provides a mechanism to combine these information together.

Furthermore, the unified approach outperforms the other model evaluation methods mentioned in Chapter 2. The test of close and not close fit method is not able to take the user-defined maximal trivial misspecification at the parameter level. The modification indices and power approach failed to provide appropriate rejection rates. As shown in Study 1, it led to high rejection rates for trivially misspecified models if sample size was low. The model evaluation methods using Bayesian analysis are sensitive to the choices of informative priors. Furthermore, the rejection rates from these methods were influenced by size of factor loadings, sample size, and numbers of items, which is not a desirable property. The performance of the simulation approach was very close to that of the unified approach. The simulation approach, however, erroneously assumes that failure to reject a model implies a well-fitting model.

In conclusion, the unified approach can solve most of the issues existing in the current practices of model evaluation. However, the unified approach is not limitless.

6.3 Limitations of the Unified Approach

6.3.1 Require Large Sample Size

The unified approach provides inconclusive results unless there is enough information. Therefore, large sample size is required unless a model is extremely severely misspecified. In Study 1, the proportion of inconclusive results was less than 50% when the sample size was higher than or equal to 500. In contrast, with the sample size of 500, the test of close fit (the null RMSEA = .05 and the alternative RMSEA = .08) and the test of not close fit (the null RMSEA = .05 and the alternative RMSEA = .01) had sufficient power to detect misspecified model examined in Study 1 (df = 20; MacCallum et al., 1996). Thus, the sample size required for the unified approach is larger than the

test of close fit and not close fit.[1] On the other hand, although the unified approach requires large sample, it provides more accurate results comparing to the other model evaluation approaches. Thus, researchers need to make sure that they have a large enough sample size to obtain conclusive results from the unified approach. A priori sample size estimation is highly recommended for the unified approach. The sample size estimation will be discussed in the extensions section. If researchers cannot obtain a sufficient sample size for the unified approach, the simulation approach is recommended. However, researchers must be aware of the problem of erroneously assuming failure to reject a model as a well-fitting model.

6.3.2 Long Computation Time

On average, the unified approach used 15 and 28 minutes to complete a single analysis on average in Studies 1 and 2, respectively. The global fit evaluation used most of the time. More specifically, the repeated sampling method and the analyses on generated data sets (with maximal trivial or minimal severe misspecifications) are two major sources for the time cost. The data analyses took much longer than the repeated sampling method. The time cost was computed with the shortcut illustrated in Steps 2a-4a and Steps 5a-7a in Chapter 3 taken into account.

To reduce the time cost, users may apply the local fit evaluation before the global fit evaluation. Based on Tables 5.6 and 5.10 in Chapter 5, the results from global and local fit evaluations did not contradict each other. Thus, rather than starting with global fit evaluation, users may start with local fit evaluation, which runs faster. If the results from the local fit evaluation were underpowered, researchers can further investigate the global fit evaluation. In other words, researchers do not need to waste their time on global fit evaluation if local fit evaluation has provided conclusive results.

Another way to save time in the global fit evaluation is to reduce the number of fit indices to be examined. For example, RMSEA, CFI, and TLI led to similar conclusions as they are highly correlated (Beauducel & Wittmann, 2005). Thus, users may drop some of them. The easy way

[1] To my knowledge, I cannot find the methods of sample size estimations for the other model evaluation methods, except using a Monte Carlo simulation.

to reduce the number of fit indices is to use collinearity analysis (Hair et al., 2006). Collinearity analysis can be applied to the resulting fit indices from the repeated sampling method in searching the maximal trivial misspecification (or minimal severe misspecification). For example, suppose 100 draws are used to find the maximal trivial misspecification, RMSEA, CFI, TLI, and SRMR from the 100 draws can be then used to calculate the tolerance for each fit index. If any fit index has a tolerance less than .10, the fit index might be dropped. Note that the fit indices should be dropped one at a time.

6.3.3 Subjectivity of the Uses of the Unified Approach

Users need to make many decisions during the procedure of the unified approach. First, they need to pick specific types of misspecification and corresponding level of maximal trivial misspecification. Second, they need to pick the alpha level, which subsequently determines the confidence level of the confidence intervals. Users also need to select the fit indices to be included in global fit evaluation. The standards for the last two decisions can be easily made. Selecting the standards for the type of misspecification and the level of maximal trivial misspecification, however, is not an easy task.

The types of misspecification may vary models. Although I have reviewed several types of misspecification (factor loadings, measurement error correlation, factor correlation, or regression coefficients), they are applicable to relatively simple models (e.g., confirmatory factor analysis). These types of misspecification are also by no mean exhaustive. For example, the differences between factor loadings within a construct is a potential source of misspecification in a tau-equivalent model, which are not included in the past research. To make the result more generalizable, it is recommended to account for sources of misspecification as many as possible. To do so, researchers may start with the list of modification indices by treating all fixed parameters as sources of misspecification. Then, they can consider other misspecifications besides the fixed parameters. For example, an extra factor with small variance may be included as a trivial misspecification in a CFA model.

For all types of misspecification, researchers need to set the level of maximal trivial misspecification. The level of maximal trivial misspecification could be different across substantive areas. As shown in Chapter 3, I reviewed several types of misspecification considered in the past research. Some studies suggested guidelines on the level of misspecification. I reviewed the range of values that are considered trivial or severe misspecifications in the previous studies. Unfortunately, there is no consensus on these values. For example, either .1 or .3 may be used as the threshold for maximal trivial misspecification for measurement error correlations. Because of the subjectivity of the decision, there is a chance that researchers will abuse it by picking the value that would lead to a preferred result. However, the problem is not unique to the unified approach but also exists in the one-size-fit-all cutoffs. For instance, researchers may pick the CFI cutoff of .90 (Bentler & Bonett, 1980) or .95 (Hu & Bentler, 1999) to get their desired results. Even though this problem from one-size-fit all cutoffs remains in the unified approach, the parameters values that researchers used to specify the level of maximal trivial misspecification are more meaningful than specific values of fit indices. For example, researchers know the amount of misfit by specifying the level of misspecification at cross loadings (e.g., not greater than .3). In contrast, researchers do not know the amount of misspecified cross loadings for a given value of a fit index because fit indices are influenced by model characteristics (e.g., the number of items or the amount of target factor loadings). Alternatively, researchers may set the level of maximal trivial misspecification by examining whether the results from misspecified models provide accurate parameter estimates. This issue is discussed in the last section of the limitations of the unified approach.

On the one hand, to set up a standard for model evaluation in structural equation modeling, a consensus needs to be reached in using the unified approach in terms of the common types of misspecification and level of maximal trivial misspecification for a specific type of model. The standard will facilitate the communication among researchers when a well-fitting or bad-fitting model is claimed. On the other hand, different substantive areas can have different levels of maximal trivial misspecification. A high-risk study may set up a stricter standard on specifying the level of maximal trivial misspecification. For example, researchers may specify the error correlations of

.2 as the maximal trivial misspecification for a scale used in correlational research. However, if a scale is used for individual assessment (e.g., intelligence test), the level of maximal trivial misspecification for error correlation can be lower (e.g., .05). Omitted error correlations of .2 may lead to slightly distorted target factor loadings and factor scores for each individual. The distorted factor scores can change the category of diagnosis (e.g., from low average to borderline in intelligence tests).

Based on the discussion above, there is not a good way to solve the subjectivity problem because a single standard cannot be applied to all substantive areas. However, to add more objectivity into the procedure, I recommend that a standard should be established for the maximal trivial misspecification in each substantive area. Once the standard is established, justification is required if a researcher proposes to use values different from the standard. This information should be transparently reported to readers. If readers disagreed with the proposed values, they may reanalyze the data with different sets of maximal trivial misspecifications. Furthermore, researchers may implement the unified approach by two or more sets of maximal trivial misspecifications. This practice will show how the conclusion is sensitive to different sets of maximal trivial misspecification.

6.3.4 Inconclusive Results

As shown above, the unified approach had a higher chance to produce an inconclusive result than the other approaches. This feature is not desirable because researchers do not know how to proceed. The proportions of inconclusive results from the unified approach are high, especially for small sample size. Across all replications in Studies 1 and 2, the proportions of inconclusive results were 49% and 62%, respectively. Researchers need to narrow down the reasons behind the inconclusive results: low sample size (no enough power) or the degree of misspecification is close to the maximal trivial misspecification. The reasons can be narrowed down using local fit evaluation. If the decision from local fit evaluation is underpowered, larger sample size is needed. Otherwise, the degree of misspecification is close to the level of maximal trivial misspecification. In the latter case, ideally, researchers should still get a larger sample size until conclusive results are obtained.

Alternatively, researchers may adjust the level of maximal trivial misspecification to be slightly higher. For example, researchers may specify the level of maximal trivial misspecification of measurement error correlations as .2 initially and then try .3. If they obtain a trivial misspecification result with the correlations of .3, the researchers will know that the problem is not sample size but the fact that the model misspecification is close to the level of trivial misspecification. However, if the change in the level of maximal trivial misspecification still provides an inconclusive result, this is more likely a small sample size problem. In this case, researchers may consider other model evaluation methods. As described above, the simulation approach is a good alternative compared to the other model evaluation methods.

6.3.5 Hidden Concerns of Well-Fitting Models from the Unified Approach

The unified approach solves the problems related to the use of one-size-fit-all fit indices cutoffs. The unified approach, however, still share some common problems with the other model fit evaluation methods when it is concluded that a model fits data well. First, the unified approach cannot differentiate between equivalent models. Equivalent models have the same values of discrepancy functions and fit indices although they are qualitatively different in interpretation (Tomarken & Waller, 2003, 2005). For example, the full mediation model from observed variables X to Y via M is equivalent to the model of Y to X via M. Both models would provide almost equal fit in the unified approach.[2] Therefore, researchers can only rely on theories or research designs to differentiate between equivalent models (Bollen, 2000; Tomarken & Waller, 2003, 2005). For example, if X and M precede Y in time, the mediation model of Y to X via M is not plausible.

Second, there exists nonequivalent models providing equal or better fit than a hypothesized model (Olsson et al., 2000; Tomarken & Waller, 2003, 2005). Thus, researchers should acknowledge that the unified approach only provides the evidence that the current hypothesized model is one of the possible explanations of the relations underlying data. Furthermore, one should try to

[2]Note that the results of equivalent models based on the unified approach may not be exactly the same because the fixed parameters of equivalent models can be based on different scales. For instance, the misspecified direct effect from X to Y is standardized by different values from the misspecified direct effect from Y to X.

rule out other possible explanations using theories, research designs, or model comparisons.

Finally, model fit evaluation is an omnibus test of the fixed parameters. Good model fit is achieved when constraints imposed on hypothesized models (e.g., fixed parameters or equality constraints) do not significantly reduce the reproducibility of model implied means and covariances to sample means and covariances. All model evaluation methods described above (including the unified approach) only test whether the constraints as a set are plausible[3]. These methods do not test whether estimated target parameters are different from 0 or have practically significant effect sizes. Thus, researchers still need to investigate the magnitude of the effect for each target parameter.

6.3.6 Symmetric Confidence Intervals of EPCs

The unified approach assumes that the sampling distribution is normal for each parameter in building confidence intervals of EPCs. This assumption may be violated when sample size is small, especially for specific parameters, such as variances or correlations. If the sampling distribution is not normal, the confidence intervals of EPCs for these parameters are not accurate which may cause erroneous results. However, local fit evaluation requires large sample size to get conclusive results so the model decision from the local fit evaluation was not biased as shown in Studies 1 and 2.

6.3.7 Accurate Parameter Estimates

All models are misspecified to some degree (MacCallum & Austin, 2000). The unified approach tries to identify a model that is close to the data-generating model such that the degree of misspecification is trivial. However, the list of parameters from a hypothesized model does not match perfectly with the list of parameters from the model used in data generation. Researchers hope that the averaged parameter estimates from the hypothesized model are not different from the corre-

[3] Although local fit evaluation can check individual constraints, the results from individual constraints are combined to evaluate global fit.

sponding true parameter values used to generate data (i.e., unbiased). Olsson et al. (2000) showed that the discrepancies in parameter estimates was negligible if maximum likelihood was used. The discrepancies decreased as sample size increased. However, it is better to check the discrepancies for the model with the maximal trivial misspecification imposed because biased parameter estimates could result in misleading research conclusions.

The unified approach allows researchers to do it during the global fit evaluation, multiple data sets are generated from the model with maximal trivial misspecification. The average parameter estimates from these data sets can be compared to the data-generating parameter values (i.e., the prameter estimates from fitting the hypothesized model from the observed data. If the discrepancies are substantial (e.g., change from small to medium effect sizes), the amount of misspecification imposed in the model is too large. Researchers should reduce their level of maximal trivial misspecification.

6.4 Extensions

This dissertation introduces the unified approach for absolute model fit with complete normally distributed data. In this section, I discuss how the unified approach can be extended to nonnormal data and data with missing observations. I also provide a sample size estimation method for the unified approach. In addition, a nested model comparison using the unified approach is introduced.

6.4.1 Nonnormally Distributed Data

I will discuss two types of nonnormally distributed data: nonnormal continuous variables and ordered categorical variables. Last two centuries, most data in psychology are treated as continuous and Micceri (1989) showed that most data sets were not normally distributed. Fortunately, maximum likelihood provides accurate parameter estimates even under nonnormality (Olsson et al., 2000). However, the chi-square values, fit indices, and the standard errors of parameter estimates will be biased by nonnormality (Yuan, 2005). Because the parameter estimates are accurate, the

model-implied mean vector and covariance matrix are still accurate. The data in global fit evaluation of the unified approach are generated based on multivariate normal distribution. Bollen and Stine's (1992) bootstrap may be used for data generation (see Equation 3.1) to account for nonnormal distribution. Local fit evaluation is based on the confidence interval of EPC. However, to my knowledge, the impacts of nonnormality on the confidence intervals of EPCs has not been evaluated.

If data are ordered categorical variables, estimation method designed specifically for categorical data can be used (Flora & Curran, 2004). One popular method assumes that each ordered categorical variable has an underlying latent normal distribution and thresholds are used to categorize continous scores into categories. To take ordinal data into account, for the global fit evaluation, multivariate normal distributed data can be generated first. Then, thresholds are applied to transform continuous variables into categorical variables. For local fit evaluation, the obtained modification indices can be used to obtain the confidence intervals of EPCs directly.

6.4.2 Missing Data

Most social science data sets have missing observations. Given that all variables predicting missing data patterns (i.e., missing at random) are included, two methods appropriately provide accurate parameter estimates and standard errors: multiple imputation (Rubin, 1987) and full information maximum likelihood (Arbuckle, 1996). See Enders (2010), Graham (2009), and Schafer & Graham (2002) for further details on missing data analysis.

6.4.2.1 Global Fit Evaluation

If data are continuous, researchers may use either multiple imputation or full information maximum likelihood to obtain parameter estimates in global fit evaluation. Then, a maximal trivial misspecification is added and model-implied mean vector and covariance matrix are computed. After that, researchers need to create data with missing observations. Savalei & Yuan (2009) proposed a modified Bollen-Stine bootstrap procedure accounting for missing data. This method can

be used to transform the data in a way so that the mean and covariance structure is equal to the model-implied mean vector and covariance matrix with maximal trivial misspecification. Their methods also retains the missing data mechanisms if missing at random is assumed. Then, the data with missing observations can be analyzed by multiple imputation or full information maximum likelihood.

If data are categorical, multiple imputation is a better option in analyzing data with missing at random (Asparouhov & Muthén, 2010). Full information maximum likelihood is not available. Reserachers cannot use the Bollen-Stine bootstrap approach because the method was designed for continous data. The method of combining bootstrap and multiple imputation proposed by Wu & Jia (2013) may be modified. First, missing values in the observed data are imputed by multiple imputation. For each imputation, global fit evaluation is implemented by the unified approach for ordered categorical variables. Then, the obtained fit indices values from all imputations are mixed together. The fit indices cutoffs can be obtained from the mixed distribution.

6.4.2.2 Local Fit Evaluation

If full information maximum likelihood is used, the obtained modification indices can be used to find the confidence intervals of EPCs directly. If multiple imputation is used, researchers need to obtain the parameter estimates and standard errors of EPCs. Then, Rubin's (1987) rule may be used to pool the EPCs and their standard errors across imputed data sets.

6.4.3 Sample Size Estimation

In structural equation modeling, sample size is usually estimated by power analysis. The main framework of power analysis in model fit evaluation is based on the test of close fit or not close fit (MacCallum et al., 1996, 2006). Sample size should be large enough so that severely misspecified model can be rejected with sufficient power for the test of close fit or the trivially misspecified model can be retained with sufficient power for the test of not close fit. Note that power analysis can be done based on the significance testing of specific parameters using a Monte Carlo simulation

(Muthén & Muthén, 2002). Another approach to estimate sample size is based on the width of the confidence interval for parameter estimates (Lai & Kelley, 2011) or RMSEA (Kelley & Lai, 2011). The confidence interval should be sufficiently narrow to achieve accurate estimates of parameters and avoid the inconclusive result (i.e., confidence interval brackets the maximal trivial misspecification value) in test of close fit / not close fit. I will show that both approaches for sample size estimation are applicable for the unified approach.

Based on the simulation studies provided above, larger sample size decreases the proportion of inconclusive results in the local fit evaluation. The global fit evaluation is less sensitive to sample size and takes long computational time so I recommend to ignore the sample size estimation for the global fit evaluation. For power analysis, the unified approach provides two types of statistical power: the power in retaining trivially misspecified models and the power in rejecting severely misspecified models. Thus, researchers should provide two models: a trivially misspecified model that researchers really wish to retain with a high rate (referred to as M_T) and a severely misspecified model that researchers really wish to reject with a high rate (referred to as M_S). For example, researchers may specify the level of maximal trivial misspecification as the measurement error correlations of .2. They may set M_T and M_S as a model with the measurement error correlations of .1 and .5, respectively. Researchers may use the repeated sampling method to find M_T or M_S (see Steps 2 and 5 of the unified approach in Chapter 3, respectively).

Then, multiple data sets with a given sample size can be created from M_T and M_S. Then, the local fit evaluation can be used to check the proportion of inconclusive results from M_T and M_S. Researchers can adjust the sample size until they achieve their desired proportions of inconclusive results, such as less than .2 for both M_T and M_S. Researchers may consider different types of misspecifications depending on their research purposes. For example, as suggested in Wu et al. (2009), misspecification in the mean structure and covariance structure should be both investigated in power analysis of growth curve model.

The local fit evaluation uses confidence intervals of EPCs to evaluate models. The width of confidence intervals should be narrow enough for two purposes: (a) to get accurate EPCs (or other

parameter estimates in a model) or (b) to lower the chance of getting the inconclusive or underpowered results. I will focus on the latter objective. The width of confidence intervals of EPCs should be narrow enough to avoid inconclusive results. Initially, researchers need to find the desired width of the confidence intervals of EPC. This can be determined by the level of maximal trivial misspecification and the specification of M_T and M_S. Then, for each fixed parameter, the differences between the maximal trivial misspecification and the values from M_T and M_S are computed. These differences represent one side of the confidence intervals of EPCs. The smallest difference is picked and the value is multiplied by 2. The result is the desired width. For example, for a measurement error correlation, the maximal trivial misspecification is .2 and the values on M_T and M_S are .1 and .5 so the differences are .1 and .3, respectively. The desired width is $.1 \times 2 = .2$. If M_T and M_S are not specified, researchers may specify the width that avoids underpowered results such that the width is less than the range of trivial misspecification. For example, if the maximal trivial misspecification is .2, the desired width is .4 or less.

A Monte Carlo simulation can be used to estimate sample size to get the expected width of the confidence intervals of EPC. The procedure is similar to power analysis described above. First, researchers generate multiple data sets from M_T and M_S. Then, these simulated data sets are are fitted by the hypothesized model. The width of the confidence intervals of EPCs are calculated. Researchers can find the average of the widths and adjust their sample size values until the desired widths of all confidence intervals of EPCs are achieved. Note that the observed width of the confidence interval is random across repeated sampling. Researchers may want to make sure that the observed width is narrower the desired width by $C\%$, which is referred to as degree of assurance (Kelley & Lai, 2011; Kelley & Rausch, 2006). To find the sample size for a certain degree of assurance, researchers find the C-th percentile value of the widths across simulated samples. Next, sample size is adjusted until the C-th percentile value of the width is equal to the desired width. Then, if researchers obtain a sample with the estimated sample size, there is $C\%$ probability that they get confidence intervals narrower than the desired width.

6.4.4 Nested Model Comparison

The problems of the current uses of fit indices for absolute model fit were described and I proposed the unified approach to solve the problems. In practice, researchers may have multiple models (e.g., from competing theories) and would like to select one model from them, which is referred to as a model selection problem. Sometimes, both absolute model fit and model selection are used at the same time. For example, for models with both measurement model and structural model, instead of testing both measurement model and structural model by absolute model fit, researchers may test the measurement model by the absolute model fit first and then select a best structural model through model comparison (Anderson & Gerbing, 1988; see Hayduk & Glaser, 2000, and Bollen, 2000, for the comments of using these steps). In this paper, I will focus on nested model comparison that is frequently used in multiple-step model evaluation, such as the two-step approach or a measurement invariance testing.

6.4.4.1 Background

If one model can be created by fixing or relaxing some parameters from the other model, two models are nested. The model that has more parameters is called the parent model and the model with fewer parameters is called the nested model. For example, a model with both measurement and structural parts is nested in a model with the measurement part only. The chi-square test statistic can be also used in selecting between two nested models. The model-implied mean vectors and covariance matrices from two nested models are compared. The chi-square statistic tests whether the discrepancy is attributed to sampling error. Similar to the chi-square test for absolute model fit, the chi-square test for nested model comparison is sensitive to sample size if the nested and parent models are trivially different.

Consequently, practical fit indices are proposed as an effect size measure of the difference between model-implied mean vectors and covariance matrices from two hypothesized models. For example, the difference in CFI is used to establish measurement invariance across groups (Cheung & Rensvold, 2002; Meade et al., 2008) by a cutoff (e.g., .002 or .01). If the difference

in CFI is lower than the cutoff, two models are equivalent and the nested model (measurement-invariant model) is preferred. As an effect size measure, the difference in CFI can be interpreted as the difference in the misfit of two models compared to the range of misfit between baseline and saturated model:

$$\Delta CFI = \frac{F_0 - F_2}{F_0 - F_S} - \frac{F_0 - F_1}{F_0 - F_S} = \frac{F_1 - F_2}{F_0 - F_S} = \frac{F_1 - F_2}{F_0} \tag{6.1}$$

where F_1 and F_2 are the discrepancy values of two models.

The fit indices in model comparison are also sensitive to the influencing factors described above, including model characteristics. For example, Meade & Bauer (2007) showed that lower factors to indicators ratio (i.e., higher factor overdetermination) and higher reliability increased the change in CFI. Therefore, a one-size-fit-all cutoff for nested model comparison is not achievable. Furthermore, most simulation studies (except Cheung & Rensvold, 2002) do not impose any parsimony error such that two nested models are negligibly different in the population. Consequently, alternative procedures for nested model comparison are needed.

6.4.4.2 Alternative Approaches for Nested Model Comparison

The test of close fit and not close fit can be generalized for nested model comparison (MacCallum et al., 2006; Li & Bentler, 2011). Researchers first specify the amount of misspecification of both nested and parent models. They must specify the levels of misspecification such that nested and parent models are equivalent (or trivially different). For example, population RMSEAs for the nested and parent models of .04 and .06 are considered equivalent. That is, the difference of .02 is trivial. Then, the sampling distribution of chi-square values can be derived from the trivial difference. Li & Bentler (2011) showed the relations between specifying trivial difference in absolute model fit and nested model comparison. The main problem of this approach is the use of population fit indices to specify maximal trivial misspecification. Population fit indices are dependent on model characteristics so they should not be used for specifying maximal trivial misspecification.

The modification indices and power approach does not have a direct extension for nested model comparison. This framework, however, can be easily applied to nested model comparison. The fixed parameters defining the differences between the nested and parent models can be investigated by the significance of the modification indices and the power to detect trivial misspecification. The same problems of using this method for absolute model fit are still applied to nested model comparison.

There are multiple frameworks in comparing models in Bayesian analysis. As the most popular framework, researchers may use Bayes factor in comparing between nested and nonnested models (Gelman et al., 2004; van de Schoot et al., 2012; Kass & Raftery, 1995). To my knowledge, the performance of the bayes factor in structural equation modeling currently has not been investigated yet. The simpler form of the bayes factor is Bayesian Information Criterion (BIC; Schwarz et al., 1978). The performance of BIC in model selection is highly influenced by sample size Preacher & Merkle (2012) such that one model is preferred in small sample sizes but another model is preferred in large sample sizes.

The simulation approach can be extended to nested model comparison as well (Pornprasertmanit et al., 2013). First, both nested and parent models are fitted to the observed data and a desired fit index (e.g., the difference in chi-square value) is saved. Then, the parameter estimates from the nested model is imposed by a set of trivial misspecification. For example, small differences in factor loadings are added to non-invariant items across groups. Then, multiple data sets are generated from the parameter estimates with trivial misspecification. The data sets are fitted by both nested and parent models and the desired fit indices are obtained. The fit index values from the simulated data are compared with the observed fit index. As the major limitation, the simulation method for the nested model comparison still erroneously assumes that failure to reject a model means a well-fitting model.

6.4.4.3 The Extension of the Unified Approach for Nested Model Comparison

The unified approach can be easily modified for nested model comparisons. For the global fit evaluation, the fit indices for nested model comparison are used instead of the fit indices designed for absolute model fit. For the local fit evaluation, the method remains the same by evaluating the absolute model fit of the nested model. However, the fixed parameters that are the differences between the nested model are emphasized.

Global Fit Evaluation. The method for global fit evaluation in nested model comparison is similar to the method for absolute model fit. Initially, the observed data are fitted by both nested and parent models. The target fit indices for model comparison are saved, such as the differences in CFI, chi-square values, Akaike Information Criterion (AIC; Akaike, 1973) and BIC. Next, the parameter values in the nested model are used for data generation. The parameter values may come from theories or previous research as well.

Next, the trivial misspecification can be imposed on the nested population model by the population misfit method, the fixed parameter method, or the repeated sampling method. For example, in weak measurement invariance testing, researchers may use the fixed parameter method by making two standardized factor loadings different in the magnitude of .1. The repeated sampling method for maximal trivial misspecification can be used. The domain of trivial misspecification is defined. Then, multiple sets of trivial misspecification are drawn and the fit indices for model comparison are calculated. The sets with the worst fit indicated by the fit indices are used for data generation. Both parent and nested models are fitted to the generated data sets and the fit indices are calculated from the results. Then, users can compare the observed fit indices with the sampling distribution of fit indices from generated data sets. If the observed fit indices indicate worse fit than most of generated fit indices, the nested model is rejected and the parent model is selected. Otherwise, the decision is still inconclusive. The shortcut (Steps 2a-4a) can be used also.

The minimal severe misspecification can be imposed as well. Researchers may randomly pick one dimension of misfit defining the difference between nested models. Each dimension is specified at the threshold of maximal trivial misspecification. For example, four factor loadings are

constrained to be equal across groups for weak invariance testing. Each factor loading is put at the thresholds, such as -.1 and .1. Thus, in this case, there are eight possible candidates for minimal severe misspecification. Then, the fit indices are used to evaluate all eight candidates and the set with the best fit indicated by the fit indices are picked. Then, data are generated from the model with minimal severe misspecification. If the observed fit indices indicate better fit than most of the generated fit indices, the nested model is selected. Otherwise, the decision is inconclusive. The shortcut (Steps 5a-7a) can be used. The results from different fit indices can be combined with the similar method as the unified approach for absolute model fit.

Local Fit Evaluation. The results from fitting the nested model to the data are used. The confidence intervals of EPCs are still used to evaluate local fits. However, the fixed parameters defining the differences between nested models are used. For example, if four factor loadings from Indicators 2 to 5 are constrained to be equal across groups for weak invariance testing (Indicator 1 is used for scale identification), researchers only focus on the confidence intervals of EPCs from the factor loadings from Indicators 2 to 5. The confidence intervals are compared with the range of trivial misspecification and identified whether the fixed parameter is severely misspecified, trivially misspecified, inconclusive, or underpowered. The results across fixed parameters can be pooled by the method used in the absolute model fit.

6.5 The Limitations of the Simulation Studies and Future Studies.

Although two simulation studies described above support the unified approach, these simulation studies are not enough to claim that the unified approach is good in every situation. There are many limitations in the simulation studies. First, only two types of models are examined: confirmatory factor analytic models and growth curve models. Future studies could investigate the performance of the unified approach for more types of models including path analysis, autoregressive model, exploratory factor analysis, growth curve analysis with different types of trajectories, and multi-

level structural equation modeling. Second, the performance of the unified approach for growth curve models was not compared to the other model evaluation methods in the second study. Thus I do not know whether the unified approach outperforms the other model evaluation methods in this scenario. Third, the influence of the level of maximal trivial misspecification on the performance of the unified method was not investigated in the second study.

Future studies may study the performance of the unified approach on all extensions described in this chapter. First, researchers may investigate the performance of using Bollen-Stine bootstrap in the unified approach to see whether it appropriately handles nonnormal continuous variables. Second, the performance of the unified approach on categorical indicators could be examined. Researchers may check the performance of the unified approach if the method for continuous data (used in this dissertation) is applied to categorical indicators. Third, the performance of the unified approach with the existence of missing data should be investigated. Finally, the performance of the unified approach for nested model comparison should be investigated as well, especially in the situations that researchers typically use nested model comparison, such as the two-step approach (measurement model vs. structural model) and measurement invariance testings.

6.6 Conclusions

Practical fit indices with one-size-fit-all cutoffs are widely used in structural equation models. In spite of the popularity, most analysts are not fully aware of the properties of practical fit indices. This dissertation provides an extensive review of the properties of practical fit indices. The review reveals several problems of the current use of the practical fit indices:

1. The performance of fit indices cutoffs depends on model characteristics, sample size, model type, and data distribution.

2. The derivation of the one-size-fit-all cutoffs is arbitrary.

3. Different fit indices may lead to inconsistent results.

Alternative methods have been proposed to account for parsimony error: the test of close fit and not close fit (Browne & Cudeck, 1992), the modification indices and power approach (Saris et al., 2009), Bayesian approach (Muthén & Asparouhov, 2012), and the simulation approach (Millsap, 2013). These approaches have their own limitations and advantages so I enhance these approaches by combining the advantages of these methods together, referred to as the unified approach.

I used two simulation studies to investigate the performance of the unified approach and compare its performance to the other model evaluation methods. The results showed that the unified approach outperformed other model evaluation methods, especially in large sample size. The unified approach retained trivially misspecified models and rejected severely misspecified models, was negligibly influenced by sample size and model characteristics, and consistently rejected severely misspecified models across all types of misspecification.

However, the unified approach has several limitations. First, large sample size (≥ 500) is required to make sure that the unified approach provides a conclusive result. Second, the global fit evaluation in the unified approach is time-consuming. Third, the unified approach is as subjective as the other model evaluation methods, especially in specifying the level of maximal trivial misspecification. Fourth, researchers may obtain inconclusive results that are not desirable as an analysis outcome. Fifth, the unified approach cannot rule out the possibility of equivalent models or nonequivalent models with better fit. Researchers must use their theories or research designs to rule out equivalent or nonequivalent models. Finally, the unified approach is designed for underspecified models but not overspecified models.

Although this dissertation only considers normally distributed variables with complete data. I showed that the unified approach can be extended to nonnormal data and missing data. It can be also extened to nested model comparison. Power analysis can be performed within the framework of the unified approach.

To fully utilize the advantages of the unified approach, I urge analysts to carefully define trivial misspecification so that appropriate cutoffs can be identified. The unified approach does not abandon the use of cutoffs for the fit indices (in global fit evaluation). Rather, it provides a framework to

identify appropriate cutoffs for fit indices. This approach encourages analysts to understand their models and specify every step in model evaluation rather than blindly using the one-size-fit-all cutoff from others' suggestions, which they have no idea where the suggestions come from (Lance et al., 2006).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & C. F (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akadéminai Kiadó.

Anderson, J. C. & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411–423.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.

Asparouhov, T. & Muthén, B. (2010). *Multiple imputation with Mplus* (tech. rep.). retrieved from http://www.statmodel.com/download/imputations7.pdf.

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

Beauducel, A. & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1), 41–75.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, 42(5), 825–829.

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.

Bentler, P. M. & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods*, 15(2), 111–123.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika*, 51(3), 375–377.

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

Bollen, K. A. (2000). Modeling strategies: In search of the holy grail. *Structural Equation Modeling*, 7(1), 74–81.

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, 21(2), 205–229.

Boulton, A. J. (2011). *Fit index sensitivity in multilevel structural equation modeling* (Unpublished master thesis). University of Kansas, Lawrence, KS.

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, 47(6), 904–930.

Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.

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.

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

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

Cheung, G. W. & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods*, 4(3), 236–264.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research*, 32(2), 208–252.

Davey, A. (2005). Issues in evaluating model fit with missing data. *Structural Equation Modeling*, 12(4), 578–597.

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

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). Boca Raton, FL: CRC press.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Fan, X. & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, 12(3), 343–367.

Fan, X. & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42(3), 509–529.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6(1), 56–83.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532–538.

Flora, D. B. & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466–491.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (in press). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, XX, XX–XX.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4), 733–760.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Hoboken, NJ: Pearson.

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.

Hayduk, L. A. & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling*, 7(1), 1–35.

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.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, 19(1), 36–50.

Holzinger, K. J. & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution. Supplementary Educational Monographs.* Chicago, IL: University of Chicago.

Hu, L.-t. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424–453.

Hu, L.-t. & 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.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling*, 14(1), 48–76.

Jones, L. V. & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5(4), 411–414.

Jöreskog, K. G. & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago, IL: Scientific Software International.

Jorgensen, T., Garnier-Villarreal, M., Lee, J., Pornprasertmanit, S., & Little, T. D. (in preparation). *Detecting omitted parameters in Bayesian confirmatory factor analysis: A Monte Carlo simulation study*. Unpublished manuscript.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kelley, K. & Lai, K. (2011). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research*, 46(1), 1–32.

Kelley, K. & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2), 137–152.

Kelley, K. & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11(4), 363–385.

Kenny, D. A. & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10(3), 333–351.

Kirkwood, T. B. & Westlake, W. J. (1981). Bioequivalence testing–a need to rethink. *Biometrics*, 37(3), 589–594.

La Du, T. J. & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, 74(4), 625–635.

Lai, K. & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, 16(2), 127–148.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods*, 9(2), 202–220.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling*, 18(4), 663–685.

Li, L. & Bentler, P. M. (2011). Quantified choice of root-mean-square errors of approximation for evaluation and power analysis of small differences between structural equation models. *Psychological Methods*, 16(2), 116–126.

Liu, S., Rovine, M. J., & Molenaar, P. C. (2012). Using fit indexes to select a covariance model for longitudinal data. *Structural Equation Modeling*, 19(4), 633–650.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13(3), 203.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, 38(1), 113–139.

MacCallum, R. C. & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51(1), 201–226.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11(1), 19–35.

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.

Mahler, C. (2011). *The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling* (Unpublished master thesis). University of British Columbia, Vancouver, Canada.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103(3), 391–410.

Marsh, H. W. & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Educational*, 64(4), 364–390.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11(3), 320–341.

Maydeu-Olivares, A. & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, 41(1), 55–64.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107(2), 247–255.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, 42(5), 859–867.

Meade, A. W. & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, 14(4), 611–635.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568–592.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, 42(5), 875–881.

Millsap, R. E. (2010). *A simulation paradigm for evaluating "approximate fit" in latent variable modeling*. Paper presented at the Conference Honoring the Scientific Contributions of Michael W. Browne, Ohio State University, Columbus, OH.

Millsap, R. E. (2013). A simulation paradigm for evaluating model fit. In M. Edwards & R. MacCallum (Eds.), *Current issues in the theory and application of latent variable models* (pp. 165–182). New York, NY: Routledge.

Millsap, R. E. & Lee, S. (October, 2008). *Approximate fit in SEM without a priori cutpoints*. Paper presented at the Annual meeting of Society of Multivariate Experimental Psychology.

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*, 19(1), 86–98.

Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences*, 42(5), 883–891.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105(3), 430–445.

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335.

Muthén, L. K. & Muthén, B. O. (1998–2013). *Mplus user's guide* (version 7) [computer software and manual]. Los Angeles, CA: Muthén and Muthén.

Muthén, L. K. & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599–620.

Nye, C. D. & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, 14(3), 548–570.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural equation modeling*, 7(4), 557–595.

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). A Monte Carlo approach for nested model comparisons in structural equation modeling. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (pp. 187–197). New York: Springer.

Preacher, K. J. & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17(1), 1–14.

R Development Core Team (2013). Comprehensive R archival network: http://www.cran.r-project.org/ R foundation for statistical computing.

Raykov, T. (2000). On sensitivity of structural equation modeling to latent relation misspecifications. *Structural Equation Modeling*, 7(4), 596–607.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, 3(4), 369–379.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, 95(452), 1143–1156.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Ryu, E. & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, 16(4), 583–601.

Saris, W. E. & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16(4), 561–582.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514.

Satorra, A. & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50(1), 83–90.

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.

Savalei, V. & Yuan, K.-H. (2009). On the model-based bootstrap with missing data: obtaining a p-value for a test of exact fit. *Multivariate Behavioral Research*, 44(6), 741–763.

Schafer, J. L. & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147–177.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Seaman, M. A. & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3(4), 403–411.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, 58(7), 935–943.

Steiger, J. H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9(2), 164–182.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42(5), 893–898.

Steiger, J. H. & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the Annual meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B. G. & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Old Tappan, NJ: Allyn & Bacon.

Taylor, A. B. (2008). *Two new methods of studying the performance of SEM fit indexes* (Doctoral dissertation). Retrieved from Dissertations and Theses database. (UMI No. 3318439).

Tomarken, A. J. & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology*, 112(4), 578–598.

Tomarken, A. J. & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31–65.

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

van de Schoot, R., Hoijtink, H., Hallquist, M. N., & Boelen, P. A. (2012). Bayesian evaluation of inequality-constrained hypotheses in SEM models using Mplus. *Structural Equation Modeling*, 19(4), 593–609.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 32, 741–744.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology*, 8, 84–136.

Widaman, K. F. & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8(1), 16–37.

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

Wu, W. (2008). *Evaluating model fit for growth curve models in SEM and MLM frameworks* (Doctoral dissertation). Retrieved from Dissertations and Theses database. (UMI No. 3339593).

Wu, W. & Jia, F. (2013). A new procedure to test mediation with missing data through nonparametric bootstrapping and multiple imputation. *Multivariate Behavioral Research*, 48(5), 663–691.

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

Wu, W. & West, S. G. (2013). Detecting misspecification in mean structures for growth curve models: Performance of pseudo R^2s and concordance correlation coefficients. *Structural Equation Modeling*, 20(3), 455–478.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, 14(3), 183–201.

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

Yuan, K.-H. & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological measurement*, 64(5), 737–757.

Yung, Y.-F. & Schumacker, R. E. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & J. S. Long (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ: Erlbaum.

Appendix A

Misspecified Models of All Designs across Simulation Studies

Table A.1 provides the population fit indices of each design across simulation studies defining misspecified models.

Table A.1: Population Fit Indices of Each Type of Misspecifications based on Sample Size of 156.

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Hu & Bentler (1999): Three Factor CFA							
Simple 1	88	19.57	.399	.047	.125	.966	.960
Simple 2	89	26.71	.569	.055	.153	.954	.946
Complex 1	85	38.12	.805	.067	.054	.950	.938
Complex 2	56	69.49	.992	.090	.067	.908	.888
Beauducel & Wittmann (2005): CFA							
Four-orthogonal-factor CFA with main LOAD of 0.4	164	5.47	.091	.018	.019	.917	.904
Four-orthogonal-factor CFA with main LOAD of 0.5	164	10.56	.146	.026	.027	.918	.905
Four-orthogonal-factor CFA with main LOAD of 0.6	164	17.37	.243	.033	.035	.919	.906
Four-orthogonal-factor CFA with main LOAD of 0.8	164	33.56	.572	.047	.051	.918	.905
Four-oblique-factor CFA with main LOAD of 0.4	164	5.35	.090	.018	.018	.921	.908
Four-oblique-factor CFA with main LOAD of 0.5	164	10.47	.145	.025	.026	.920	.908
Four-oblique-factor CFA with main LOAD of 0.6	164	17.39	.243	.033	.035	.920	.907
Four-oblique-factor CFA with main LOAD of 0.8	164	36.03	.521	.047	.051	.918	.905
Eight-orthogonal-factor CFA with main LOAD of 0.4	712	10.94	.089	.013	.013	.917	.909
Eight-orthogonal-factor CFA with main LOAD of 0.5	712	21.12	.140	.017	.019	.918	.910
Eight-orthogonal-factor CFA with main LOAD of 0.6	712	34.74	.233	.022	.025	.918	.911
Eight-orthogonal-factor CFA with main LOAD of 0.8	712	71.11	.569	.032	.037	.918	.910
Eight-oblique-factor CFA with main LOAD of 0.4	712	10.71	.088	.012	.013	.921	.913
Eight-oblique-factor CFA with main LOAD of 0.5	712	20.94	.139	.017	.019	.920	.913
Eight-oblique-factor CFA with main LOAD of 0.6	712	34.78	.234	.022	.025	.920	.912
Eight-oblique-factor CFA with main LOAD of 0.8	712	72.05	.578	.032	.036	.918	.910
Curran et al. (2003): Full SEM							
Model 1 with Level-1 Misspecification	23	2.24	.098	.031	.015	.994	.991
Model 1 with Level-2 Misspecification	24	4.45	.160	.043	.022	.988	.982
Model 1 with Level-3 Misspecification	25	10.16	.367	.064	.035	.973	.961
Model 2 with Level-1 Misspecification	86	4.34	.097	.027	.017	.994	.992
Model 2 with Level-2 Misspecification	87	8.74	.164	.032	.024	.987	.985
Model 2 with Level-3 Misspecification	88	14.5	.280	.041	.032	.979	.975
Model 3 with Level-1 Misspecification	53	12.45	.311	.045	.030	.971	.961
Model 3 with Level-2 Misspecification	54	37.25	.889	.084	.073	.913	.884
Model 3 with Level-3 Misspecification	57	52.23	.979	.096	.080	.879	.847
Fan et al. (1999): Four-factor CFA: CFA							
Level-1 Misspecification	46	27.61	.768	.078	.027	.980	.971
Level-2 Misspecification	48	96.37	1.000	.142	.044	.929	.902

Continued on next page

Table A.1 – *Continued from previous page*

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Hancock & Mueller (2011): Six-Factor Full SEM							
Standardized LOAD of .40	129	5.060441	.094	.020	.019	.939	.928
Standardized LOAD of .45	129	8.003053	.128	.025	.025	.936	.924
Standardized LOAD of .50	129	12.013748	.185	.031	.032	.932	.919
Standardized LOAD of .55	129	17.266766	.276	.037	.041	.929	.916
Standardized LOAD of .60	129	23.965399	.410	.043	.052	.927	.913
Standardized LOAD of .65	129	32.423161	.584	.050	.065	.925	.911
Standardized LOAD of .70	129	43.097218	.769	.058	.079	.924	.910
Standardized LOAD of .75	129	56.477454	.911	.067	.096	.923	.909
Standardized LOAD of .80	129	72.656636	.980	.075	.119	.924	.910
Standardized LOAD of .85	129	91.548133	.997	.085	.144	.927	.914
Standardized LOAD of .90	129	113.713527	1.000	.094	.172	.933	.921
Standardized LOAD of .95	129	140.426061	1.000	.105	.202	.944	.933
La Du & Tanaka (1989): Path Analysis							
MIS Model	29	7.01	.230	.050	.025	.980	.967
Jackson (2007): CFA							
Two-factor CFA with main LOAD of 0.4 and cross LOAD of 0.1	169	1.690576	.061	.010	.011	.990	.989
Two-factor CFA with main LOAD of 0.6 and cross LOAD of 0.1	169	3.717362	.076	.015	.018	.994	.993
Two-factor CFA with main LOAD of 0.8 and cross LOAD of 0.1	169	9.1055	.127	.023	.026	.994	.994
Two-factor CFA with main LOAD of 0.4 and cross LOAD of 0.2	169	5.587874	.092	.018	.018	.974	.971
Two-factor CFA with main LOAD of 0.6 and cross LOAD of 0.2	169	14.789786	.200	.030	.033	.978	.975
Two-factor CFA with main LOAD of 0.8 and cross LOAD of 0.2	169	44.621442	.713	.052	.054	.976	.973
Two-factor CFA with main LOAD of 0.4 and cross LOAD of 0.3	169	9.717569	.134	.024	.022	.965	.961
Two-factor CFA with main LOAD of 0.6 and cross LOAD of 0.3	169	32.418607	.506	.044	.045	.960	.955
Two-factor CFA with main LOAD of 0.8 and cross LOAD of 0.3	169	137.397147	1.000	.091	.099	.949	.942
Two-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.1	169	0.3006557	.052	.004	.004	.998	.998
Two-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.1	169	0.5805767	.054	.006	.007	.999	.999
Two-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.1	169	1.488786	.059	.009	.009	.999	.999
Two-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.2	169	1.7550613	.061	.010	.010	.991	.990
Two-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.2	169	3.6972019	.076	.015	.015	.994	.994
Two-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.2	169	13.2392984	.178	.028	.021	.993	.992
Two-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.3	169	5.1537511	.088	.018	.016	.979	.977
Two-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.3	169	12.5861821	.169	.027	.024	.983	.981
Two-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.3	169	65.1118648	.922	.062	.047	.974	.971
Five-factor CFA with main LOAD of 0.4 and cross LOAD of 0.1	160	0.9147464	.056	.008	.006	.993	.991
Five-factor CFA with main LOAD of 0.6 and cross LOAD of 0.1	160	2.5551264	.067	.013	.011	.994	.993

Continued on next page

Table A.1 – *Continued from previous page*

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Five-factor CFA with main LOAD of 0.8 and cross LOAD of 0.1	160	7.8513458	.116	.022	.016	.994	.993
Five-factor CFA with main LOAD of 0.4 and cross LOAD of 0.2	160	3.5771165	.075	.015	.012	.977	.973
Five-factor CFA with main LOAD of 0.6 and cross LOAD of 0.2	160	11.0923347	.154	.026	.020	.978	.974
Five-factor CFA with main LOAD of 0.8 and cross LOAD of 0.2	160	44.8907421	.733	.053	.033	.969	.964
Five-factor CFA with main LOAD of 0.4 and cross LOAD of 0.3	160	8.2684726	.121	.023	.017	.960	.952
Five-factor CFA with main LOAD of 0.6 and cross LOAD of 0.3	160	28.9472527	.456	.043	.029	.953	.944
Five-factor CFA with main LOAD of 0.8 and cross LOAD of 0.3	160	276.3343664	1.000	.132	.055	.865	.840
Five-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.1	160	0.1597631	.051	.003	.003	.999	.998
Five-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.1	160	0.428826	.053	.005	.005	.999	.999
Five-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.1	160	1.2732993	.058	.009	.008	.999	.999
Five-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.2	160	0.7106621	.054	.007	.006	.995	.994
Five-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.2	160	2.021073	.063	.011	.010	.996	.995
Five-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.2	160	7.1543536	.109	.021	.017	.995	.994
Five-factor CFA with main LOAD of 0.4 with one MIS factor with LOAD of 0.3	160	1.7497862	.062	.011	.008	.991	.989
Five-factor CFA with main LOAD of 0.6 with one MIS factor with LOAD of 0.3	160	5.5057981	.092	.019	.015	.991	.989
Five-factor CFA with main LOAD of 0.8 with one MIS factor with LOAD of 0.3	160	24.560447	.373	.039	.030	.988	.985
Heene et al. (2012): Two-Factor CFA							
Low main LOAD with three MIS positive error covariances	251	207.38	1.000	.091	.005	.989	.988
Low main LOAD with six MIS positive error covariances	251	462.17	1.000	.136	.008	.975	.973
Low main LOAD with three MIS positive and negative error covariances	251	46.76	.622	.043	.003	.997	.997
Low main LOAD with six MIS positive and negative error covariances	251	84.18	.950	.058	.004	.995	.995
High main LOAD with three MIS positive error covariances	251	211.41	1.000	.092	.002	.994	.993
High main LOAD with six MIS positive error covariances	251	169.5	1.000	.138	.002	.986	.985
High main LOAD with three MIS positive and negative error covariances	251	48.09	.641	.044	.001	.999	.998
High main LOAD with six MIS positive and negative error covariances	251	86.6	.958	.059	.001	.997	.997
Heene et al. (2011): CFA							
Five items for each factor; Simple misspecification; Low main LOAD	87	8.273	.156	.031	.044	.893	.871
Five items for each factor; Simple misspecification; Moderate main LOAD	87	18.0161	.364	.046	.088	.921	.905
Five items for each factor; Simple misspecification; High main LOAD	87	25.9269	.557	.055	.131	.940	.927
Five items for each factor; Complex misspecification; Low main LOAD	90	28.3479	.603	.056	.073	.635	.574
Five items for each factor; Complex misspecification; Moderate main LOAD	90	69.0091	.990	.088	.136	.698	.647
Five items for each factor; Complex misspecification; High main LOAD	90	114.4287	1.000	.113	.192	.734	.689
Fifteen items for each factor; Simple misspecification; Low main LOAD	936	22.2152	.135	.016	.052	.948	.945
Fifteen items for each factor; Simple misspecification; Moderate main LOAD	936	33.074	.190	.019	.098	.970	.968
Fifteen items for each factor; Simple misspecification; High main LOAD	936	38.2698	.223	.020	.140	.980	.979
Fifteen items for each factor; Complex misspecification; Low main LOAD	945	99.2771	.708	.033	.078	.777	.767

Continued on next page

Table A.1 – *Continued from previous page*

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Fifteen items for each factor; Complex misspecification; Moderate main LOAD	945	201.504	.995	.046	.142	.818	.809
Fifteen items for each factor; Complex misspecification; High main LOAD	945	323.2294	1.000	.059	.199	.831	.823
Nye & Drasgow (2011): Two-Factor Categorical CFA (Fit measures are calculated from polychoric COR)							
Simple misspecification	89	12.2708	.230	.037	.042	.966	.960
Complex misspecification	90	27.1494	.576	.055	.123	.926	.913
Schermelleh-Engel et al. (2003): Four-Factor Full SEM							
Mild misspecification	15	5.2906	.234	.060	.055	.986	.973
Severe misspecification	16	22.0843	.865	.118	.119	.940	.896
Taylor (2008): Three-Factor CFA							
Main LOAD of 0.5; Factor COR of 0.2; One cross-LOAD of 0.2	51	1.953585	.075	.020	.017	.980	.975
Main LOAD of 0.8; Factor COR of 0.2; One cross-LOAD of 0.2	51	7.059935	.173	.037	.029	.989	.985
Main LOAD of 0.5; Factor COR of 0.5; One cross-LOAD of 0.2	51	0.9964837	.062	.014	.011	.992	.989
Main LOAD of 0.8; Factor COR of 0.5; One cross-LOAD of 0.2	51	4.874811	.126	.031	.020	.993	.991
Main LOAD of 0.5; Factor COR of 0.2; One cross-LOAD of 0.5	51	9.161199	.225	.043	.034	.923	.900
Main LOAD of 0.8; Factor COR of 0.2; One cross-LOAD of 0.5	51	57.30128	.992	.107	.069	.916	.891
Main LOAD of 0.5; Factor COR of 0.5; One cross-LOAD of 0.5	51	4.215811	.113	.029	.021	.971	.963
Main LOAD of 0.8; Factor COR of 0.5; One cross-LOAD of 0.5	51	34.98455	.871	.083	.051	.953	.939
Main LOAD of 0.5; Factor COR of 0.2; Two cross-LOAD of 0.2	51	3.916488	.108	.028	.024	.963	.952
Main LOAD of 0.8; Factor COR of 0.2; Two cross-LOAD of 0.2	51	14.14672	.368	.053	.041	.977	.971
Main LOAD of 0.5; Factor COR of 0.5; Two cross-LOAD of 0.2	51	2.004002	.076	.020	.015	.985	.980
Main LOAD of 0.8; Factor COR of 0.5; Two cross-LOAD of 0.2	51	9.769928	.241	.044	.029	.986	.982
Main LOAD of 0.5; Factor COR of 0.2; Two cross-LOAD of 0.5	51	20.00356	.545	.063	.050	.862	.821
Main LOAD of 0.8; Factor COR of 0.2; Two cross-LOAD of 0.5	51	117.8826	1.000	.153	.098	.845	.799
Main LOAD of 0.5; Factor COR of 0.5; Two cross-LOAD of 0.5	51	9.387212	.231	.043	.030	.949	.934
Main LOAD of 0.8; Factor COR of 0.5; Two cross-LOAD of 0.5	51	72.62787	.999	.120	.068	.914	.888
Main LOAD of 0.5; Factor COR of 0.2; Three cross-LOAD of 0.2	51	5.910476	.147	.034	.030	.947	.931
Main LOAD of 0.8; Factor COR of 0.2; Three cross-LOAD of 0.2	51	21.2631	.582	.065	.050	.967	.957
Main LOAD of 0.5; Factor COR of 0.5; Three cross-LOAD of 0.2	51	3.046986	.092	.025	.018	.979	.972
Main LOAD of 0.8; Factor COR of 0.5; Three cross-LOAD of 0.2	51	14.70488	.385	.054	.035	.979	.973
Main LOAD of 0.5; Factor COR of 0.2; Three cross-LOAD of 0.5	51	34.72027	.868	.083	.063	.797	.737
Main LOAD of 0.8; Factor COR of 0.2; Three cross-LOAD of 0.5	51	180.6453	1.000	.189	.120	.785	.722
Main LOAD of 0.5; Factor COR of 0.5; Three cross-LOAD of 0.5	51	16.94828	.454	.058	.037	.925	.903
Main LOAD of 0.8; Factor COR of 0.5; Three cross-LOAD of 0.5	51	121.0253	1.000	.155	.082	.872	.834
Wu (2008): Growth Curve Model							
Model 1: Mildly misspecify variance of quadratic factor	8	0.2031484	.057	.016	.005	1.000	1.000

Continued on next page

Table A.1 – *Continued from previous page*

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Model 2: Moderately misspecify variance of quadratic factor	8	0.322264	.061	.020	.007	.999	.999
Model 3: Severely misspecify variance of quadratic factor	8	1.4762081	.107	.043	.016	.997	.996
Model 4: Mildly misspecify covariance between intercept and linear factors	8	1.7107816	.117	.046	.058	.997	.996
Model 5: Moderately misspecify covariance between intercept and linear factors	8	2.7790671	.170	.059	.074	.995	.993
Model 6: Severely misspecify covariance between intercept and linear factors	8	13.128104	.732	.129	.126	.974	.968
Model 7: Mildly misspecify equality constraint at error variances	8	1.3186572	.100	.041	.010	.998	.997
Model 8: Moderately misspecify equality constraint at error variances	8	1.9588558	.129	.050	.014	.996	.995
Model 9: Severely misspecify equality constraint at error variances	8	3.5633528	.213	.067	.026	.993	.991
Model 10: Mildly misspecify first-order autoregressive regression among errors	8	0.1003155	.053	.011	.003	1.000	1.000
Model 11: Moderately misspecify first-order autoregressive regression among errors	8	0.1598207	.055	.014	.004	1.000	1.000
Model 12: Severely misspecify first-order autoregressive regression among errors	8	0.7055613	.075	.030	.009	.999	.998
Model 13: Mildly misspecify the mean of quadratic factor	8	1.6365821	.114	.045	.011	.997	.996
Model 14: Moderately misspecify the mean of quadratic factor	8	2.6461878	.163	.058	.014	.995	.994
Model 15: Severely misspecify the mean of quadratic factor	8	12.539958	.708	.126	.031	.976	.970
Model 16: Model 13 + Model 1	9	1.846117	.119	.046	.012	.996	.996
Model 17: Model 14 + Model 1	9	2.874356	.167	.057	.015	.994	.994
Model 18: Model 15 + Model 1	9	12.967423	.703	.121	.032	.975	.972
Model 19: Model 13 + Model 2	9	1.994044	.125	.047	.013	.996	.996
Model 20: Model 14 + Model 2	9	3.031416	.175	.058	.016	.994	.993
Model 21: Model 15 + Model 2	9	13.198339	.713	.122	.033	.974	.971
Model 22: Model 13 + Model 3	9	3.21665	.184	.060	.020	.993	.993
Model 23: Model 14 + Model 3	9	4.298255	.242	.069	.023	.991	.990
Model 24: Model 15 + Model 3	9	14.933804	.777	.129	.039	.970	.966
Model 25: Model 13 + Model 4	9	3.35095	.191	.061	.059	.994	.993
Model 26: Model 14 + Model 4	9	4.364676	.245	.070	.060	.992	.991
Model 27: Model 15 + Model 4	9	14.273242	.754	.127	.065	.972	.969
Model 28: Model 13 + Model 5	9	4.411626	.248	.070	.075	.991	.990
Model 29: Model 14 + Model 5	9	5.419199	.305	.078	.075	.989	.988
Model 30: Model 15 + Model 5	9	15.280592	.788	.131	.079	.970	.967
Model 31: Model 13 + Model 6	9	14.477842	.761	.127	.117	.972	.969
Model 32: Model 14 + Model 6	9	15.299395	.789	.131	.113	.970	.967
Model 33: Model 15 + Model 6	9	23.234412	.946	.161	.098	.955	.950
Model 34: Model 13 + Model 7	9	2.982786	.172	.058	.014	.994	.994
Model 35: Model 14 + Model 7	9	4.006891	.226	.067	.017	.992	.992
Model 36: Model 15 + Model 7	9	14.022747	.745	.125	.031	.974	.971
Model 37: Model 13 + Model 8	9	3.593218	.204	.064	.018	.993	.992

Continued on next page

Table A.1 – *Continued from previous page*

Conditions	*df*	*ncp*	power	RMSEA	SRMR	CFI	TLI
Model 38: Model 14 + Model 8	9	4.60261	.259	.072	.019	.991	.990
Model 39: Model 15 + Model 8	9	14.477585	.761	.127	.033	.973	.970
Model 40: Model 13 + Model 9	9	5.099892	.287	.076	.028	.990	.989
Model 41: Model 14 + Model 9	9	6.052842	.342	.082	.029	.988	.986
Model 42: Model 15 + Model 9	9	15.426061	.793	.132	.041	.969	.965
Model 43: Model 13 + Model 10	9	1.76466	.115	.045	.011	.997	.996
Model 44: Model 14 + Model 10	9	2.795509	.163	.056	.014	.994	.994
Model 45: Model 15 + Model 10	9	12.875906	.699	.120	.032	.975	.972
Model 46: Model 13 + Model 11	9	1.823062	.118	.045	.012	.996	.996
Model 47: Model 14 + Model 11	9	2.853016	.166	.057	.015	.994	.994
Model 48: Model 15 + Model 11	9	12.925546	.701	.120	.033	.975	.972
Model 49: Model 13 + Model 12	9	2.361218	.142	.051	.014	.995	.995
Model 50: Model 14 + Model 12	9	3.386425	.193	.062	.017	.994	.993
Model 51: Model 15 + Model 12	9	13.41055	.721	.123	.035	.974	.971
Fan & Sivo (2007): CFA or Full SEM							
Three-factor CFA with Level-1 MIS cross LOAD	85	22.2032	.473	.051	.033	.972	.966
Three-factor CFA with Level-2 MIS cross LOAD	86	44.7567	.886	.073	.056	.944	.932
Three-factor CFA with Level-1 MIS factor COR	88	79.8295	.998	.096	.265	.903	.884
Three-factor CFA with Level-2 MIS factor COR	89	101.0914	1.000	.107	.312	.877	.855
Four-factor SEM with Level-1 MIS cross LOAD and factor residual COR	48	18.4972	.516	.062	.021	.986	.980
Four-factor SEM with Level-2 MIS cross LOAD and factor residual COR	49	35.2148	.882	.085	.034	.973	.963
Two-factor CFA with Level-1 MIS cross LOAD and factor COR	1	0	.050	.000	.000	1.000	1.000
Two-factor CFA with Level-2 MIS cross LOAD and factor COR	2	5.0867	.511	.160	.030	.973	.920
Three-factor SEM with Level-1 MIS structural path and measurement error COR	5	7.1869	.513	.121	.045	.971	.914
Three-factor SEM with Level-2 MIS structural path and measurement error COR	8	18.1301	.881	.151	.062	.928	.864
Davey (2005): Three-Factor CFA							
Misspecify at the measurement model; factor LOAD = 0.4; factor COR = 0.4)	24	2.8472	.113	.035	.025	.894	.842
Misspecify at the measurement model; factor LOAD = 0.8; factor COR = 0.4)	24	57.7303	1.000	.156	.113	.853	.779
Misspecify at the measurement model; factor LOAD = 0.4; factor COR = 0.8)	24	0.3915	.057	.013	.008	.991	.987
Misspecify at the measurement model; factor LOAD = 0.8; factor COR = 0.8)	24	18.4636	.689	.088	.035	.963	.945
Misspecify at the structural model; factor LOAD = 0.4; factor COR = 0.4)	25	2.117	.093	.029	.029	.922	.887
Misspecify at the structural model; factor LOAD = 0.8; factor COR = 0.4)	25	11.9228	.438	.069	.122	.970	.956
Misspecify at the structural model; factor LOAD = 0.4; factor COR = 0.8)	25	8.7541	.311	.060	.060	.800	.712
Misspecify at the structural model; factor LOAD = 0.8; factor COR = 0.8)	25	59.8814	1.000	.156	.270	.880	.827

Note. df = degree of freedom. ncp = noncentrality parameter. RMSEA = Root mean square error of approximation. SRMR = Standardized root mean squared residuals. CFI = Comparative fit index. TLI = Tucker-Lewis index. LOAD = Loadings. COR = Correlations. MIS = Misspecified.

Appendix B

Misspecified Values for Maximal Trivial Misspecifications

Table B.1 shows trivial parameter values providing the maximal trivial misspecifications for RMSEA, CFI, and TLI. Table B.2 shows trivial parameter values providing the maximal trivial misspecifications for RMSEA, CFI, and SRMR.

Table B.1: Parameter Values Proving the Maximal Trivial Misspecification for RMSEA, CFI, and TLI

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Standardized Factor Loadings																			
Spatial	**.814**	**.392**	**.441**	**.578**	-.181	-.140	-.175	.050	.061	-.024	-.123	.097	.059	.089	.092	.115	-.049	.089	-.075
Verbal	.089	-.004	-.062	-.055	**.836**	**.806**	**.870**	**.724**	**.838**	-.167	.172	.005	.083	.147	.137	-.031	-.105	-.079	.101
Speed	.063	.102	-.098	-.036	-.162	.162	-.033	-.004	.026	**.557**	**.800**	**.517**	**.537**	.199	.127	.066	-.060	.094	.131
Memory	-.081	.070	-.023	.063	.147	-.159	-.066	.196	-.183	-.071	.108	-.180	.001	**.628**	**.526**	**.596**	**.561**	**.472**	**.467**
Ignored Factor	.175	.149	.081	.125	-.120	.035	.155	-.152	-.155	-.044	.173	.027	.150	-.067	-.036	-.082	.153	.006	-.177
Measurement Error Correlations																			
1. Visual perception	1																		
2. Cubes	.050	1																	
3. Paper from board	.188	.162	1																
4. Flags	-.102	-.144	-.014	1															
5. General information	-.093	.114	-.097	-.069	1														
6. Paragraph comprehension	.015	-.153	.137	.187	-.132	1													
7. Sentence completion	.023	.028	.035	-.090	-.026	.181	1												
8. Word classification	-.128	-.072	.122	.121	.005	-.158	-.052	1											
9. Word meaning	-.041	.151	.093	-.156	-.124	.099	.044	.102	1										
10. Addition	.078	.198	-.141	.019	.113	.042	.044	.123	.122	1									
11. Code	-.100	-.077	.067	-.079	-.018	-.152	-.002	.116	-.180	-.150	1								
12. Counting groups of dots	.069	.189	.069	-.097	-.146	.064	.081	-.039	.118	.069	.033	1							
13. Straight and curved capitals	-.055	-.135	.049	-.013	-.161	-.085	-.142	-.093	-.088	.126	.179	.090	1						
14. Word recognition	.194	-.133	.012	-.110	.083	.044	-.178	-.068	-.178	-.117	-.097	.170	.123	1					
15. Number recognition	-.104	-.165	-.061	-.177	.160	.005	.067	.168	-.068	.024	.198	-.062	.090	-.182	1				
16. Figure recognition	.117	-.135	-.025	.124	.163	-.024	.001	.009	.051	-.188	-.070	.130	.061	-.064	-.073	1			
17. Object-number	-.056	-.128	-.087	.156	-.190	.106	.118	-.153	.176	.060	-.128	-.191	-.159	.173	.103	.152	1		
18. Number-figure	-.095	.114	.068	.085	-.095	.145	.132	-.028	-.089	-.115	-.085	-.075	-.081	-.161	.010	-.008	.067	1	
19. Figure-word	-.163	.062	-.072	.116	.044	-.159	.109	-.051	.177	.154	.113	-.170	-.110	-.165	.056	.167	-.140	.063	1

Note. The boldface numbers represent standardized factor loadings of target parameters.

Table B.2: Parameter Values Proving the Maximal Trivial Misspecification for SRMR

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Standardized Factor Loadings																			
Spatial	**.814**	**.392**	**.441**	**.578**	.167	-.108	.183	-.026	.137	-.069	-.158	-.003	.059	.142	-.053	-.148	-.036	.198	-.037
Verbal	.140	-.037	-.130	.054	**.836**	**.806**	**.870**	**.724**	**.838**	-.052	-.167	-.138	.191	-.080	-.178	.072	-.177	.135	.138
Speed	.052	.149	-.038	-.126	.013	.178	-.186	.200	-.136	**.557**	**.800**	**.517**	**.537**	.089	-.149	-.194	.016	.051	.200
Memory	.018	.108	.017	-.174	.133	-.033	.116	-.014	-.155	.170	-.025	.076	.155	**.628**	**.526**	**.596**	**.561**	**.472**	**.467**
Ignored Factor	-.040	.167	.187	-.139	.085	-.079	.179	-.189	-.153	.182	-.083	-.101	-.181	.002	-.001	.139	.088	.016	.076
Measurement Error Correlations																			
1. Visual perception	1																		
2. Cubes	.054	1																	
3. Paper from board	.055	.080	1																
4. Flags	.119	.127	.132	1															
5. General information	-.116	.101	.026	-.133	1														
6. Paragraph comprehension	-.009	.192	-.131	.005	.128	1													
7. Sentence completion	.143	-.019	-.030	.172	-.183	-.087	1												
8. Word classification	-.154	.062	-.101	.197	.096	-.019	-.094	1											
9. Word meaning	-.103	.086	.181	.182	-.074	.124	.118	.170	1										
10. Addition	-.009	-.137	.155	-.053	-.083	-.056	-.132	.170	.147	1									
11. Code	-.095	-.002	-.114	-.200	.116	-.071	-.035	.103	-.104	-.066	1								
12. Counting groups of dots	.065	.166	-.154	-.046	-.161	-.006	-.066	.130	-.195	-.075	.137	1							
13. Straight and curved capitals	.132	.096	-.133	.120	-.180	.012	.110	.068	-.145	-.054	.177	.036	1						
14. Word recognition	-.107	.018	.019	-.018	-.138	.175	-.113	-.142	-.039	.186	.192	.059	-.129	1					
15. Number recognition	.198	-.081	.149	-.004	-.105	-.138	.093	.033	-.061	.005	.022	.146	.060	-.025	1				
16. Figure recognition	.191	.165	.101	-.154	-.137	-.009	-.038	.133	.119	.171	-.150	.055	.006	-.009	.149	1			
17. Object-number	-.028	-.098	-.086	.167	-.163	-.129	-.019	-.118	-.173	-.158	-.104	-.039	.085	-.010	.108	-.035	1		
18. Number-figure	.077	-.075	-.134	-.101	.039	.099	.103	-.146	-.113	.094	.099	.029	.113	-.129	-.152	.117	-.117	1	
19. Figure-word	-.041	-.025	.045	-.094	-.021	-.139	-.089	.150	.149	-.181	.164	.059	-.053	-.152	-.142	.136	.052	.058	1

Note. The boldface numbers represent standardized factor loadings of target parameters.

Appendix C

Confidence Intervals of Expected Parameter Changes

Table C.1 shows the confidence intervals of expected parameter changes and the decisions based on the unified approach.

Table C.1: Confidence Intervals of Expected Parameter Changes (EPC) and the Results of Local Fit Evaluation on Each EPC.

Fixed Parameters	Std EPC	Lower Bound	Upper Bound	Width	Decision
Spatial = General information	-.145	-.256	.222	-.034	Inconclusive
Spatial = Paragraph comprehension	.103	-.013	.233	.220	Inconclusive
Spatial = Sentence completion	-.181	-.286	.210	-.075	Inconclusive
Spatial = Word classification	.145	.015	.260	.275	Inconclusive
Spatial = Word meaning	.170	.060	.221	.281	Inconclusive
Spatial = Addition	-.296	-.452	.313	-.140	Inconclusive
Spatial = Code	.066	-.104	.340	.236	Inconclusive
Spatial = Counting groups of dots	-.010	-.167	.315	.148	Trivial
Spatial = Straight and curved capitals	.235	.078	.314	.392	Inconclusive
Spatial = Word recognition	-.145	-.299	.309	.009	Inconclusive
Spatial = Number recognition	-.090	-.247	.314	.067	Inconclusive
Spatial = Figure recognition	.348	.193	.310	.502	Inconclusive
Spatial = Object-number	-.180	-.336	.312	-.024	Inconclusive
Spatial = Number-figure	-.009	-.168	.318	.150	Trivial
Spatial = Figure-word	.078	-.081	.318	.237	Inconclusive
Verbal = Visual perception	.399	.201	.396	.597	Severe
Verbal = Cubes	-.101	-.259	.316	.057	Inconclusive
Verbal = Paper from board	-.037	-.194	.314	.120	Trivial
Verbal = Flags	-.266	-.425	.319	-.106	Inconclusive
Verbal = Addition	-.067	-.223	.311	.088	Inconclusive
Verbal = Code	.251	.076	.349	.425	Inconclusive
Verbal = Counting groups of dots	-.106	-.262	.312	.051	Inconclusive
Verbal = Straight and curved capitals	-.127	-.283	.312	.029	Inconclusive
Verbal = Word recognition	-.075	-.203	.255	.052	Inconclusive
Verbal = Number recognition	-.236	-.366	.260	-.106	Inconclusive
Verbal = Figure recognition	.151	.023	.256	.279	Inconclusive
Verbal = Object-number	.027	-.102	.258	.156	Trivial
Verbal = Number-figure	.063	-.069	.264	.195	Trivial

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters	Std EPC	Lower Bound	Upper Bound	Width	Decision
Verbal = Figure-word	.078	-.054	.264	.210	Inconclusive
Speed = Visual perception	.144	-.049	.386	.337	Inconclusive
Speed = Cubes	-.110	-.273	.325	.052	Inconclusive
Speed = Paper from board	-.172	-.333	.322	-.011	Inconclusive
Speed = Flags	.063	-.097	.321	.223	Inconclusive
Speed = General information	-.034	-.146	.225	.079	Trivial
Speed = Paragraph comprehension	.034	-.084	.236	.152	Trivial
Speed = Sentence completion	-.067	-.174	.213	.039	Trivial
Speed = Word classification	.108	-.024	.264	.240	Inconclusive
Speed = Word meaning	.013	-.099	.224	.125	Trivial
Speed = Word recognition	-.283	-.451	.337	-.114	Inconclusive
Speed = Number recognition	-.360	-.530	.341	-.189	Inconclusive
Speed = Figure recognition	.224	.056	.337	.393	Inconclusive
Speed = Object-number	.232	.062	.339	.401	Inconclusive
Speed = Number-figure	-.011	-.183	.344	.161	Trivial
Speed = Figure-word	.237	.064	.345	.409	Inconclusive
Memory = Visual perception	.002	-.187	.379	.192	Trivial
Memory = Cubes	-.068	-.233	.329	.096	Inconclusive
Memory = Paper from board	-.269	-.432	.326	-.106	Inconclusive
Memory = Flags	.240	.080	.320	.400	Inconclusive
Memory = General information	-.148	-.246	.195	-.051	Inconclusive
Memory = Paragraph comprehension	.096	-.006	.204	.198	Trivial
Memory = Sentence completion	-.106	-.198	.184	-.014	Trivial
Memory = Word classification	.149	.035	.228	.263	Inconclusive
Memory = Word meaning	.093	-.004	.194	.190	Trivial
Memory = Addition	-.045	-.218	.345	.127	Inconclusive
Memory = Code	.073	-.115	.377	.261	Inconclusive
Memory = Counting groups of dots	-.093	-.267	.348	.080	Inconclusive
Memory = Straight and curved capitals	.045	-.129	.346	.218	Inconclusive

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Visual perception	Cubes	-.304	-.565	.521	-.043	Underpowered
Visual perception	Paper from board	.009	-.274	.565	.291	Underpowered
Visual perception	Flags	-.348	-.751	.807	.055	Underpowered
Visual perception	General information	.007	-.195	.404	.209	Underpowered
Visual perception	Paragraph comprehension	.220	.023	.394	.417	Inconclusive
Visual perception	Sentence completion	-.114	-.326	.423	.097	Underpowered
Visual perception	Word classification	.069	-.121	.379	.258	Inconclusive
Visual perception	Word meaning	.055	-.148	.405	.257	Underpowered
Visual perception	Addition	-.102	-.294	.384	.090	Inconclusive
Visual perception	Code	.017	-.231	.497	.266	Underpowered
Visual perception	Counting groups of dots	-.069	-.258	.379	.121	Inconclusive
Visual perception	Straight and curved capitals	.140	-.051	.381	.330	Inconclusive
Visual perception	Word recognition	-.021	-.224	.405	.181	Underpowered
Visual perception	Number recognition	.081	-.111	.385	.273	Inconclusive
Visual perception	Figure recognition	.105	-.093	.397	.304	Inconclusive
Visual perception	Object-number	-.169	-.364	.390	.026	Inconclusive
Visual perception	Number-figure	.111	-.078	.378	.300	Inconclusive
Visual perception	Figure-word	-.105	-.294	.377	.084	Inconclusive
Cubes	Paper from board	.150	.005	.289	.294	Inconclusive
Cubes	Flags	.206	.049	.315	.364	Inconclusive
Cubes	General information	-.147	-.297	.300	.003	Inconclusive
Cubes	Paragraph comprehension	-.117	-.263	.293	.030	Inconclusive
Cubes	Sentence completion	.038	-.118	.313	.194	Trivial
Cubes	Word classification	.100	-.042	.282	.241	Inconclusive
Cubes	Word meaning	.084	-.067	.300	.234	Inconclusive
Cubes	Addition	-.216	-.358	.284	-.074	Inconclusive
Cubes	Code	.018	-.156	.347	.192	Trivial
Cubes	Counting groups of dots	-.011	-.151	.281	.130	Trivial
Cubes	Straight and curved capitals	.134	-.007	.282	.275	Inconclusive

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Cubes	Word recognition	.003	-.145	.296	.151	Trivial
Cubes	Number recognition	.009	-.133	.284	.151	Trivial
Cubes	Figure recognition	.107	-.039	.291	.252	Inconclusive
Cubes	Object-number	-.032	-.175	.287	.112	Trivial
Cubes	Number-figure	-.179	-.319	.280	-.039	Inconclusive
Cubes	Figure-word	.006	-.134	.280	.145	Trivial
Paper from board	Flags	.053	-.110	.327	.217	Inconclusive
Paper from board	General information	.067	-.084	.302	.218	Inconclusive
Paper from board	Paragraph comprehension	.004	-.143	.295	.152	Trivial
Paper from board	Sentence completion	-.090	-.248	.315	.067	Inconclusive
Paper from board	Word classification	-.103	-.245	.285	.040	Inconclusive
Paper from board	Word meaning	.100	-.051	.303	.251	Inconclusive
Paper from board	Addition	.003	-.140	.286	.146	Trivial
Paper from board	Code	-.191	-.366	.350	-.016	Inconclusive
Paper from board	Counting groups of dots	.190	.048	.283	.331	Inconclusive
Paper from board	Straight and curved capitals	-.021	-.163	.285	.121	Trivial
Paper from board	Word recognition	-.111	-.260	.298	.038	Inconclusive
Paper from board	Number recognition	-.132	-.275	.286	.011	Inconclusive
Paper from board	Figure recognition	.013	-.134	.293	.160	Trivial
Paper from board	Object-number	-.022	-.167	.289	.122	Trivial
Paper from board	Number-figure	-.041	-.182	.282	.100	Trivial
Paper from board	Figure-word	.007	-.134	.282	.148	Trivial
Flags	General information	-.159	-.315	.313	-.002	Inconclusive
Flags	Paragraph comprehension	-.051	-.204	.306	.102	Inconclusive
Flags	Sentence completion	-.155	-.319	.327	.008	Inconclusive
Flags	Word classification	.084	-.064	.295	.231	Inconclusive
Flags	Word meaning	.105	-.052	.314	.261	Inconclusive
Flags	Addition	-.127	-.275	.296	.021	Inconclusive
Flags	Code	.056	-.126	.365	.239	Inconclusive

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Flags	Counting groups of dots	.025	-.122	.293	.172	Trivial
Flags	Straight and curved capitals	.169	.021	.295	.316	Inconclusive
Flags	Word recognition	-.037	-.192	.309	.118	Trivial
Flags	Number recognition	.083	-.065	.296	.231	Inconclusive
Flags	Figure recognition	.191	.039	.304	.343	Inconclusive
Flags	Object-number	-.078	-.228	.300	.072	Inconclusive
Flags	Number-figure	-.048	-.194	.292	.098	Trivial
Flags	Figure-word	.152	.006	.292	.298	Inconclusive
General information	Paragraph comprehension	-.132	-.313	.362	.049	Inconclusive
General information	Sentence completion	-.062	-.275	.427	.152	Underpowered
General information	Word classification	.020	-.146	.331	.185	Trivial
General information	Word meaning	.233	.040	.388	.427	Inconclusive
General information	Addition	.145	-.009	.310	.300	Inconclusive
General information	Code	.003	-.187	.380	.193	Trivial
General information	Counting groups of dots	.051	-.102	.306	.204	Inconclusive
General information	Straight and curved capitals	-.076	-.230	.308	.078	Inconclusive
General information	Word recognition	-.177	-.338	.322	-.016	Inconclusive
General information	Number recognition	.045	-.109	.309	.199	Trivial
General information	Figure recognition	.061	-.097	.317	.220	Inconclusive
General information	Object-number	-.092	-.248	.312	.064	Inconclusive
General information	Number-figure	-.070	-.223	.305	.082	Inconclusive
General information	Figure-word	-.082	-.234	.304	.070	Inconclusive
Paragraph comprehension	Sentence completion	.141	-.057	.397	.340	Inconclusive
Paragraph comprehension	Word classification	.009	-.150	.317	.167	Trivial
Paragraph comprehension	Word meaning	-.058	-.240	.364	.124	Inconclusive
Paragraph comprehension	Addition	.170	.019	.302	.321	Inconclusive
Paragraph comprehension	Code	-.048	-.233	.370	.137	Inconclusive
Paragraph comprehension	Counting groups of dots	-.087	-.236	.299	.063	Inconclusive
Paragraph comprehension	Straight and curved capitals	-.076	-.226	.300	.074	Inconclusive

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Paragraph comprehension	Word recognition	.178	.021	.314	.335	Inconclusive
Paragraph comprehension	Number recognition	-.060	-.211	.302	.090	Inconclusive
Paragraph comprehension	Figure recognition	-.061	-.216	.309	.093	Inconclusive
Paragraph comprehension	Object-number	.027	-.125	.305	.180	Trivial
Paragraph comprehension	Number-figure	.064	-.085	.298	.212	Inconclusive
Paragraph comprehension	Figure-word	.014	-.134	.297	.163	Trivial
Sentence completion	Word classification	.114	-.064	.355	.292	Inconclusive
Sentence completion	Word meaning	-.054	-.269	.430	.161	Underpowered
Sentence completion	Addition	-.249	-.411	.323	-.088	Inconclusive
Sentence completion	Code	.265	.066	.398	.463	Inconclusive
Sentence completion	Counting groups of dots	-.134	-.293	.320	.026	Inconclusive
Sentence completion	Straight and curved capitals	.012	-.148	.321	.173	Trivial
Sentence completion	Word recognition	.154	-.014	.336	.322	Inconclusive
Sentence completion	Number recognition	-.060	-.221	.322	.101	Inconclusive
Sentence completion	Figure recognition	-.183	-.348	.330	-.018	Inconclusive
Sentence completion	Object-number	.067	-.096	.326	.230	Inconclusive
Sentence completion	Number-figure	-.095	-.253	.318	.064	Inconclusive
Sentence completion	Figure-word	-.065	-.224	.317	.094	Inconclusive
Word classification	Word meaning	-.217	-.383	.332	-.051	Inconclusive
Word classification	Addition	-.041	-.187	.291	.105	Trivial
Word classification	Code	.215	.037	.356	.393	Inconclusive
Word classification	Counting groups of dots	-.075	-.220	.288	.069	Inconclusive
Word classification	Straight and curved capitals	-.127	-.272	.290	.017	Inconclusive
Word classification	Word recognition	-.055	-.206	.303	.097	Inconclusive
Word classification	Number recognition	.037	-.108	.291	.183	Trivial
Word classification	Figure recognition	.118	-.032	.298	.267	Inconclusive
Word classification	Object-number	-.018	-.165	.294	.129	Trivial
Word classification	Number-figure	.088	-.055	.287	.232	Inconclusive
Word classification	Figure-word	.046	-.097	.287	.190	Trivial

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Word meaning	Addition	.018	-.137	.310	.173	Trivial
Word meaning	Code	-.238	-.428	.381	-.047	Inconclusive
Word meaning	Counting groups of dots	.136	-.017	.307	.290	Inconclusive
Word meaning	Straight and curved capitals	.066	-.088	.308	.220	Inconclusive
Word meaning	Word recognition	-.075	-.236	.322	.086	Inconclusive
Word meaning	Number recognition	-.101	-.256	.310	.054	Inconclusive
Word meaning	Figure recognition	.113	-.046	.317	.271	Inconclusive
Word meaning	Object-number	.026	-.131	.313	.182	Trivial
Word meaning	Number-figure	.123	-.030	.305	.275	Inconclusive
Word meaning	Figure-word	.101	-.051	.305	.254	Inconclusive
Addition	Code	-.063	-.334	.543	.209	Underpowered
Addition	Counting groups of dots	.232	.074	.316	.390	Inconclusive
Addition	Straight and curved capitals	-.062	-.222	.321	.099	Inconclusive
Addition	Word recognition	-.014	-.167	.306	.139	Trivial
Addition	Number recognition	.021	-.126	.294	.168	Trivial
Addition	Figure recognition	-.072	-.223	.302	.079	Inconclusive
Addition	Object-number	.205	.057	.297	.354	Inconclusive
Addition	Number-figure	.009	-.136	.289	.153	Trivial
Addition	Figure-word	-.106	-.251	.289	.038	Inconclusive
Code	Counting groups of dots	-.313	-.564	.502	-.062	Underpowered
Code	Straight and curved capitals	-.056	-.316	.521	.205	Underpowered
Code	Word recognition	-.081	-.272	.382	.111	Inconclusive
Code	Number recognition	-.152	-.334	.363	.029	Inconclusive
Code	Figure recognition	.071	-.116	.375	.259	Inconclusive
Code	Object-number	.123	-.061	.368	.307	Inconclusive
Code	Number-figure	.015	-.164	.356	.193	Trivial
Code	Figure-word	.165	-.013	.356	.343	Inconclusive
Counting groups of dots	Straight and curved capitals	.170	.014	.311	.326	Inconclusive
Counting groups of dots	Word recognition	-.219	-.370	.303	-.067	Inconclusive

Continued on next page

Table C.1 – *Continued from previous page*

Fixed Parameters		Std EPC	Lower Bound	Upper Bound	Width	Decision
Counting groups of dots	Number recognition	-.050	-.195	.290	.095	Trivial
Counting groups of dots	Figure recognition	.007	-.142	.298	.156	Trivial
Counting groups of dots	Object-number	.024	-.123	.294	.171	Trivial
Counting groups of dots	Number-figure	.005	-.138	.286	.148	Trivial
Counting groups of dots	Figure-word	.139	-.004	.286	.282	Inconclusive
Straight and curved capitals	Word recognition	-.034	-.186	.305	.118	Trivial
Straight and curved capitals	Number recognition	-.080	-.226	.292	.066	Inconclusive
Straight and curved capitals	Figure recognition	.141	-.009	.300	.291	Inconclusive
Straight and curved capitals	Object-number	-.044	-.192	.295	.104	Trivial
Straight and curved capitals	Number-figure	-.119	-.263	.288	.025	Inconclusive
Straight and curved capitals	Figure-word	.036	-.107	.287	.180	Trivial
Word recognition	Number recognition	.241	.067	.348	.415	Inconclusive
Word recognition	Figure recognition	.015	-.173	.376	.203	Inconclusive
Word recognition	Object-number	.098	-.082	.360	.278	Inconclusive
Word recognition	Number-figure	-.090	-.257	.334	.077	Inconclusive
Word recognition	Figure-word	.016	-.150	.333	.183	Trivial
Number recognition	Figure recognition	.141	-.027	.336	.309	Inconclusive
Number recognition	Object-number	-.056	-.219	.325	.107	Inconclusive
Number recognition	Number-figure	.102	-.052	.307	.255	Inconclusive
Number recognition	Figure-word	-.262	-.415	.306	-.109	Inconclusive
Figure recognition	Object-number	-.200	-.373	.347	-.027	Inconclusive
Figure recognition	Number-figure	-.220	-.382	.323	-.059	Inconclusive
Figure recognition	Figure-word	-.059	-.220	.323	.103	Inconclusive
Object-number	Number-figure	.122	-.036	.314	.279	Inconclusive
Object-number	Figure-word	-.016	-.172	.314	.141	Trivial
Number-figure	Figure-word	.165	.016	.298	.314	Inconclusive

Note. Std = Standardized, "= " means "is loaded by". " " means "is correlated with".

Appendix D

Detailed Simulation Conditions for Simulation Study 1

The combinations of misspecifications for the 16-indicator model that are different from the 8-indicator model are described as follows:

5. **Type B Misspecification, Trivial**: The target model omits the following standardized factor loadings: $\lambda_{9,1} = .1$, $\lambda_{10,1} = .1$, $\lambda_{11,1} = .1$, and $\lambda_{12,1} = .1$.

6. **Type B Misspecification, Severe**: The target model omits the following standardized factor loadings: $\lambda_{9,1} = .3$, $\lambda_{10,1} = .3$, $\lambda_{11,1} = .3$, and $\lambda_{12,1} = .3$.

7. **Type B Misspecification, Very Severe**: The target model omits the following *unstandardized* factor loadings: $\lambda_{9,1} = 0.9$, $\lambda_{10,1} = 0.9$, $\lambda_{11,1} = 0.9$, and $\lambda_{12,1} = 0.9$.

8. **Type C Misspecification, Trivial**: The target model omits the following measurement error correlations: $\theta_{1,9} = .1$, $\theta_{2,10} = -.1$, $\theta_{3,11} = .1$, and $\theta_{4,12} = -.1$.

9. **Type C Misspecification, Severe**: The target model omits the following measurement error correlations: $\theta_{1,9} = .3$, $\theta_{2,10} = -.31$, $\theta_{3,11} = .3$, and $\theta_{4,12} = -.3$.

10. **Type C Misspecification, Very Severe**: The target model omits the following measurement error correlations: $\theta_{1,9} = .9$, $\theta_{2,10} = -.9$, $\theta_{3,11} = .9$, and $\theta_{4,12} = -.9$.

The alternative models in the simulation approach for the 16-indicator model that will be different from the 8-indicator model are described as follows:

3. The following fixed standardized cross factor loadings are changed: $\lambda_{9,1} = .1$ (or .3), $\lambda_{10,1} = .1$ (or .3), $\lambda_{11,1} = .1$ (or .3), and $\lambda_{12,1} = .1$ (or .3).

4. The following measurement error correlations are changed: $\theta_{1,9} = .1$ (or .3), $\theta_{2,10} = -.1$ (or -.3), $\theta_{3,11} = .1$ (or .3), and $\theta_{4,12} = -.1$ (or -.3).

Appendix E

Supplemental Results for Simulation Study 1

Because the unified approach can provide inconclusive results, the rejection rates from the unified approach were calculated based on conclusive results only. In contrast, other model evaluation methods calculated rejection rates based on all replications. In this appendix, I show additional results that the rejection rates for other model evaluation methods are calculated based on the conclusive replications from the unified approach. The rejection rates of all model evaluation methods are treated as missing values if the proportions of inconclusive results in any design conditions are over .90. As a result, the rejection rates for all model evaluation methods are calculated based on the same replications.

E.1 Rejection Rate for Model Misspecification and Level of Trivial Misspecification

As shown in Table E.1, four methods had nonnegligible η^2s of the interaction between degree of model misspecification and level of maximal trivial misspecification: the modification indices and power approach, the PPP method in Bayesian analysis with informative priors on cross loadings,

the simulation approach, and the unified approach. All of the other methods had rejection rates sensitive to the main effect of the degree of model misspecification or the level of trivial misspecification but not to the interaction effect.

Table E.2 provides the rejection rates classified by the degree of model misspecification and the level of trivial misspecification. The rejection rates for the trivial misspecification conditions are expected to be close to 0. As mentioned in Chapter 4, model evaluation methods with rejection rates less than .1 are to be labelled as correctly retaining models. All model evaluation methods based on the Bayesian approach had the rejection rates over .1 for at least one trivial misspecification condition. Other methods had rejection rates less than .1 for all trivial misspecification conditions.

Ideal rejection rates for the severe misspecification conditions are close to 1. The model evaluation methods with rejection rates of .9 or higher are deemed correctly rejected models. The modification indices and power approach, the simulation approach, and the unified approach provided the appropriate rejection rates in all severe misspecification conditions. Other methods had a rejection rate lower than .9 for at least one severe misspecification condition. Finally, the rejection rates for the cutoff conditions are expected to range between .1 and .9. The TLI cutoff, the combination of fit indices cutoffs, and the combination of the PPP cutoff and the zero coverage in both cross loadings and error covariances information priors provided the undesirable results such that the rejection rates were over .90 at least one cutoff condition.

Table E.1: The η^2s of the Effects of the Design Conditions on the Rejection Rates for Study 1 Selecting Only Replications that the Unified Approach Provided Conclusive Results

Factors	RMSEA	CFI	TLI	SRMR	OVCUT	CLOSE	MIPOW	Bayesian Analysis				SIM	UNIFIED
								LOAD, PPP	LOAD, ZERO	ERR, PPP	ERR, ZERO		
TYPEMIS	.000	.004	.008	**.032**	.000	.009	.001	**.311**	**.093**	**.050**	**.150**	.003	.001
N	.024	.023	.017	**.044**	.018	.021	.011	.008	.003	**.046**	.009	.002	.011
LOAD	.000	.017	.006	.010	.000	.004	.003	.000	.004	.019	.024	.000	.003
SEVERE	**.755**	**.753**	**.751**	**.679**	**.740**	**.715**	**.612**	**.164**	**.445**	**.185**	**.087**	**.670**	**.612**
LEVELMIS	.004	.002	.000	.000	.007	.000	**.100**	**.117**	**.081**	**.251**	**.122**	**.056**	**.100**
ITEMS	**.032**	.003	.002	.003	.001	.006	.004	.000	.000	**.076**	**.039**	.002	.004
TYPEMIS : N	.000	.000	.000	.000	.000	.000	.000	.020	.003	.001	.011	.000	.000
TYPEMIS : LOAD	.000	.002	.000	.011	.010	.001	.000	.000	.005	.001	.014	.000	.000
TYPEMIS : SEVERE	.000	.002	.002	.000	.000	.002	.005	.001	.008	.012	.000	.007	.005
TYPEMIS : LEVELMIS	.003	.004	.005	.003	.001	.006	.000	.014	.024	.000	**.094**	.002	.000
TYPEMIS : ITEMS	.001	.000	.001	.002	.000	.005	.002	.007	.008	.005	.000	.004	.002
N : LOAD	.000	.001	.000	.001	.000	.000	.003	.000	.000	.000	.000	.000	.003
N : SEVERE	.000	.000	.001	.000	.002	.001	.000	.019	.013	.001	.000	.001	.000
N : LEVELMIS	.000	.000	.000	.004	.000	.000	.005	.009	.004	.000	.005	.009	.005
N : ITEMS	.001	.000	.000	.000	.002	.002	.000	.000	.000	.000	.001	.000	.000
LOAD : SEVERE	.000	.001	.001	.005	.000	.001	.000	.001	.002	.011	.000	.000	.000
LOAD : LEVELMIS	.001	.002	.000	.010	.003	.004	.007	.001	.010	.016	.012	.001	.007
LOAD : ITEMS	.000	.003	.000	.000	.000	.000	.000	.000	.001	.000	.003	.000	.000
SEVERE : LEVELMIS	.002	.000	.002	.000	.016	.001	**.128**	**.051**	.025	.010	.009	**.105**	**.128**
SEVERE : ITEMS	.002	.001	.003	.000	.003	.002	.000	.005	.002	.005	**.084**	.000	.000
LEVELMIS : ITEMS	.008	.002	.007	.000	.007	.008	.000	.000	.000	.002	.005	.002	.000

Note.

1. The boldface numbers represent the η^2s $\geq .03$.

2. All interactions higher than two ways are not presented here because their η^2s $< .03$ except the three-way interaction between TYPEMIS : SEVERE : LEVELMIS on the rejection rates of LOAD, PPP (.054) and the three-way interaction between SEVERE : LEVELMIS : ITEMS on the rejection rates of ERR, PPP (.036).

3. The abbrevations for the model evaluation methods are provided at the beginning of Chapter 5.

4. TYPEMIS = Type of misspecifications, N = Sample size, LOAD = The size of target factor loadings, SEVERE = The degree of misspecification, LEVELMIS = The level of trivial misspecification, ITEMS = The number of items

Table E.2: The Rejection Rates for Each Model Evaluation Method Classified by the Level of Maximal Trivial Misspecification and the Degree of Misspecification for Study 1 Selecting Only Replications that the Unified Approach Provided Conclusive Results

Level of Trivial Misspecification	Level 1				Level 2			
Degree of Model Misspecification	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3
Classification	Trivial	Cutoff	Severe	Severe	Trivial	Trivial	Cutoff	Severe
The Proportions of Inconclusive Results								
UNIFIED	.840	.953	.462	.001	.343	.356	.874	.102
Rejection Rates								
RMSEA	.000	.000	.346	.999	.000	.000	.498	.999
CFI	.000	.000	.510	1.000	.000	.005	.829	1.000
TLI	.000	.001	.667	1.000	.002	.017	.951	1.000
SRMR	.000	.000	.490	.981	.000	.000	.489	.979
OVCUT	.000	.001	.868	1.000	.002	.017	.982	1.000
CLOSE	.000	.000	.666	1.000	.000	.000	.864	1.000
MIPOW	.000	.740	1.000	1.000	.000	.000	.661	1.000
Bayesian, LOAD, PPP	.003	.496	.309	.904	.004	.145	.343	.334
Bayesian, LOAD, ZERO	.018	.728	.897	.985	.010	.171	.978	.845
Bayesian, ERR, PPP	.250	.462	.550	.972	.043	.043	.135	.356
Bayesian, ERR, ZERO	.480	.971	.963	1.000	.315	.569	.999	.797
SIM	.000	.734	.947	1.000	.000	.001	.984	1.000
UNIFIED	.000	.740	1.000	1.000	.000	.000	.661	1.000

Note. The abbrevations for the model evaluation methods are provided at the beginning of this chapter.

E.2 The Effect of Types of Misspecification

As shown in Table E.1, All methods except the SRMR cutoff and all model evaluation methods using the Bayesian approach had negligible η^2s for all main and interaction effects involving types of misspecification. The SRMR cutoff had lower rejection rates when the misspecification was in error correlations (.48) but higher rejection rates when the misspecification was in factor correlation (.69) or cross loadings (.70).

The PPP method with cross loadings priors had nonnegligible η^2 for the interaction effect between types of misspecification, degree of misspecification, and the level of trivial misspecification. As shown in Table E.3, in the misspecifications in factor correlation or cross loadings, the rejection rates were high only when the degree of misspecification was Level 3 and the level of trivial misspecification was Level 1. However, in the misspecification in error correlations, the rejection rates were all high except when the degree of model misspecification was Level 0. The rate of increase in rejection rates was stronger for the Level 1 trivial misspecification. The PPP method with zero coverage of the credible intervals from cross loadings informative priors had

Table E.3: The Rejection Rates from the PPP Method with Cross Loadings Priors Classified by the Level of Maximal Trivial Misspecification, the Degree of Misspecification, and the Type of Misspecification for Study 1 Selecting Only Replications that the Unified Approach Provided Conclusive Results

Type of Misspecification	Factor Correlations		Cross Loadings		Error Correlations	
Level of Trivial Misspecification	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2
Degree of Model Misspecification						
Level 0	.003	.004	.003	.004	.003	.004
Level 1	.000	.004	.003	.004	.779	.426
Level 2	.012	NA	.087	.004	1.000	.935
Level 3	.742	.004	.970	.013	1.000	1.000

Note. NA is the condition that the proportion of inconclusive results from the unified approach was over .90.

lower rejection rates when the misspecification was in factor correlation (.58) but higher rejection rates when the misspecification was in cross loadings (.82) or error correlations (.87).

The PPP method with error correlations priors had rejection rates of .32, .46, and .60 for the misspecifications in factor correlation, cross loadings, and error correlations, respectively. The PPP method with the zero coverage of the credible intervals from error correlations informative priors had nonnegligible interaction effect of the type of misspecification and the level of trivial misspecification. In general, the rejection rates were high (.95 - 1.00) regardless of type of misspecification when the level of trivial misspecification was Level 1. However, when the level of trivial misspecification was Level 2, the rejection rate was lower for the misspecificaiton in factor correlation (.37) than the one in cross loadings (.88) or error correlations (.90).

E.3 The Effect of Model Characteristics

As shown in Table E.1, the size of factor loadings had negligible effects on all model evaluation methods. However, the number of items had nonnegligible effects on the RMSEA cutoff approach and the Bayesian approach using informative priors on error covariances (both ERR, PPP and ERR, ZERO). For the RMSEA cutoff approach, the rejection rate was lower when the number of items increased (rejection rates were .69 and .47 for 8 and 16 items, respectively). The rejection rates for

Table E.4: The Rejection Rates from the PPP Method with Error Covariances Priors Classified by the Level of Maximal Trivial Misspecification, the Degree of Misspecification, and the Number of Items for Study 1 Selecting Only Replications that the Unified Approach Provided Conclusive Results

The Number of Items	8		16	
Level of Trivial Misspecification	Level 1	Level 2	Level 1	Level 2
Degree of Model Misspecification				
Level 0	.000	.000	.500	.077
Level 1	.000	.000	.509	.078
Level 2	.191	.000	.793	.197
Level 3	.944	.037	1.000	.631

the PPP method had nonnegligible three-way interaction between the degree of model misspecification, the level of trivial misspecification, and the number of items. As shown in Table E.4, in the eight-item model, the rejection rate was high only when the degree of model misspecification was Level 3 and the level of trivial misspecification was Level 1. However, in the sixteen-item model, the rejection rates increased if the level of trivial misspecification was Level 1 or the degree of model misspecification was increased.

When the PPP method and the zero coverage of nontarget credible intervals are combined, the interaction effect between the degree of model misspecification and the number of items was nonnegligible. The rejection rates between eight- and sixteen-item models were not different if the degree of model misspecification was Level 2 or Level 3 (.89 - .99). However, if the degree of model misspecification was Level 0 or Level 1, the rejection rates from eight-item model (.06 and .31 for Levels 0 and 1 degree of misspecification) were lower than one from sixteen-item model (.58 and .84 for Level 0 and 1 degree of misspecification).

E.4 The Effect of Sample Size

Only the SRMR cutoff and the PPP method with error covariances informative priors had nonnegligible η^2s for the main effect of sample size. There was no nonnegligible interaction effect

involving sample size in any model evaluation methods. When sample size increased, both SRMR cutoff provided lower rejection rates (.95 and .43 for sample size of 125 and 4000, respectively), as well as the PPP method (.87 and .32 for sample size of 125 and 4000, respectively).

In sum, if only the replications that the unified approach provided conclusive results were considered, the modification indices and power approach, the simulation approach, and the unified approach satisfied all of the desired properties. As shown in Table E.1, the effect size of the interaction between the degree of misspecification and the level of trivial misspecification was non-neglibible for all three approaches. As shown in Table E.2, appropriate rejection rates were provided for trivial, severe, or cutoff conditions. Sample size, type of misspecification, number of items, and size of factor loadings did not influence the rejection rates from all three approaches. However, the desired properties of the modification indices and power approach and the simulation approach relied on the fact that the unified approach provided conclusive results. If the unified approach provided inconclusive results, the results from the two approaches are not trustworthy.