

Evaluating Reasonableness Tests for Longitudinal Measurement Invariance in Structural Equation Modeling using Confirmatory Factor Analysis

By

Elizabeth M. Grandfield

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.

Co-Chair: Holger Brandt

Co-Chair: Emily Cramer

Amber Watts

Michael S. Vitevitch

James R. Thompson

Date Defended: May 7, 2019

The dissertation committee for Elizabeth M. Grandfield certifies that this
is the approved version of the following dissertation:

**Evaluating Reasonableness Tests for Longitudinal Measurement
Invariance in Structural Equation Modeling using Confirmatory
Factor Analysis**

Co-Chair: Holger Brandt

Co-Chair: Emily Cramer

Date Approved: May 7, 2019

Abstract

Researchers are typically interested in comparing groups of people and/or comparing people across time. If researchers are to conclude differences are due to group dynamics or time, we must establish that the measure(s) we are using are actually invariant across groups or time. Some statistical methods (ANOVA and regression) make this assumption without direct evaluation. Conducting analyses in the Structural Equation Modeling (SEM) framework using Confirmatory Factor Analysis (CFA) is one way the assumption of measurement invariance can be evaluated directly. Many researchers have studied multiple group invariance and current invariance testing recommendations are based on multiple group studies and simulations. There is a lack of literature on testing invariance in longitudinal designs. Current guidelines recommend researchers apply the same guidelines from multiple group to longitudinal designs. Longitudinal designs are more complicated and may need different recommendations. The current study evaluates measurement invariance in longitudinal CFA in order to ascertain if the current guidelines based off the multiple group case are acceptable when applied to the longitudinal framework.

Acknowledgments

There are many people who I owe a great deal of gratitude to for their support and wisdom throughout this process. I would like to start by thanking each of my committee members—Drs. Emily Cramer, Holger Brandt, Amber Watts, Jim (James) Thompson, and Michael Vitevitch—for taking the time to serve on my committee, providing their expertise, guidance, and support.

I would also like to thank Drs. Todd Little, Katherine Masyn, and Audrey Leroux for their continued wisdom, advice, and generous readiness to sequester me away to provide uninterrupted writing retreats whether in an office or at Chateau Leroux. Without you I might still be staring at my cursor on a blank screen, or making sure my inbox is at zero, or fending off 1 million pecking baby geese.

Thank you to my cohort and colleagues Drs. Mauricio Garnier Villarreal, Terrence Jorgensen, and Alexander Schoemann who took time to answer my simulation questions, without your support I might still be coding. I am so thankful for your detailed responses, patience, and clarity. And to my writing group Team Awesome, thank you for your honesty, shoulder, and support. Tagging in for our writing blocks has helped us create a solid writing habit that will continue throughout our careers.

Finally, I would like to thank my family for supporting me while I moved half way across the country to pursue my goal at the University of Kansas and also forgiving my antisocial behavior while I finished my dissertation. I could not have done this without you. I especially want to thank Scott Grandfield! Thank you for your support, encouragement, and willingness to explore wherever life takes us. I can't wait to see what's next!

Table of Contents

Chapter 1: Introduction	1
Chapter 2: MI in Longitudinal SEM.....	3
Structural Equation Modeling Framework	4
Introduction to MI.....	5
Steps of Evaluating MI in CFA	6
Configural Invariance.	6
Weak Invariance model.	11
Strong Invariance.	12
Strict Invariance.	13
Special Considerations in Longitudinal CFA and MI Evaluation	14
Current Limitations.....	17
Simulation Studies	19
Purpose of the Current Study.....	20
Chapter 3: Method	21
Data Generations.....	21
Invariant models.....	21
Variant model simulation conditions.	22
Implementation of data generation.	25
Data Analysis Models.....	25
Alternative null model.	25
Measurement invariance analysis models.....	25
Implementation in R.	26

Outcomes	27
Chapter 4: Results	28
Type I Error Rates of the scaled $\Delta\chi^2$ and $\Delta AFI's$	28
Power of the scaled $\Delta\chi^2$ and $\Delta AFI's$	29
Small LMI in Factor Loadings	30
Small LMI in Item Intercepts	33
Large LMI in Item Intercepts	33
Chapter 5: Discussion	36
Research Questions	36
Performance of Chi-square difference	36
Performance of the Change in AFI's	36
Appropriate Null Model	38
Number of Occasions	38
Limitations	39
Future Directions	40
Conclusion	40
References	42

List of Figures

Figure 1. Single Latent Construct Example	15
Figure 2. Two Repeated Measurement Occasions.....	16
Figure 3. Invariant Model with Two Occasions	22
Figure 4. Small Loading LMI	23
Figure 5. Small Intercept LMI	24
Figure 6. Change in CFI for Loadings	31
Figure 7. Small Intercept CFI Change	35

List of Tables

Table 1. Manipulated Variables in Monte Carlo Design	22
Table 2. Type I Error Rates.....	29
Table 3. Power to Detect LMI in Loadings	32
Table 4. Power to Detect LMI in Intercepts	34

Chapter 1: Introduction

Factor analysis has a long history in the social and behavioral sciences. A review of research studies shows numerous versions and types of surveys, questionnaires, and scales used for measuring a variety of different constructs. Factor analysis has been an essential statistical tool for scale development and validation (Brown, 2015). Specifically, confirmatory factor analysis (CFA), in which factor structures are hypothesized a priori, in the structural equation modeling (SEM) framework has advanced and become more commonly used in the past few decades. Brown (2015) and Kline (2016) suggest that the popularity and advancements are due, at least in part, to the increased availability and improvement of technology and computing power. These improvements have been coupled with the availability of computer software developed specifically to analyze more advanced statistical techniques and provided researchers with the improved resources to analyze data using these techniques.

SEM is a latent, sometimes referred to as an unobserved, variable modeling technique that is an extension of regression and factor analysis (Bollen, 1989). Latent variables (also called latent constructs) are measured by manifest variables; such as items on a scale or questionnaire. For example, personality questions on a survey attempt to tap into a latent construct such as extraversion. An individual's level of extraversion is said to give rise to their responses on the items. SEM has many benefits as an analytic tool such as: less restrictive and testable assumptions compared to classical statistics (e.g. analysis of variance), ability to account for measurement error, flexibility, and the ability to evaluate measurement invariance between groups and/or across time (Brown, 2015; Kline, 2016; Little, 2013). The current study focuses on this last benefit of SEM: measurement invariance (MI) testing in the confirmatory factor analysis (CFA) framework with an emphasis on longitudinal CFA.

The purpose of this study is to consider the theoretical and practical issues involved in MI evaluation of longitudinal data using the currently available guidelines that were developed and tested using multiple group cross-sectional studies. Additionally, the present research will provide some background on SEM as it pertains to evaluating MI, including current popular guidelines and cutoff values implemented in MI testing based on the current literature. I will also point out gaps in the literature in regard to utilizing current cut-off criteria in the longitudinal CFA framework along with a proposed simulation study to begin testing the tenability of these guidelines with longitudinal data.

The organization of the paper is as follows: First, an overview of key MI concepts using longitudinal CFA in an SEM framework will be discussed. Specifically, I will introduce SEM concepts that are related to MI and MI evaluation. For a more detailed review of SEM please see Kline (2016) and Bollen (1989) as seminal works. This overview will include the most current and common steps and guidelines employed in the evaluation of MI. Next, gaps in the literature regarding how to apply MI evaluation in longitudinal CFA designs will be identified. Chapter 3 will outline the methodology and simulation models used to answer the key research questions. Chapter 4 outlines the results found regarding the appropriateness of utilizing cross-sectional MI guidelines to longitudinal CFAs. Finally, Chapter 5 will elaborate on the implications of these findings, discuss limitations and directions for future research.

Chapter 2: MI in Longitudinal SEM

Our world, society, and people undergo constant change and development. In order to begin to understand and explain these influences, longitudinal studies are a vital research tool that allows investigators to follow participants over time in order to collect data over repeated measurement occasions. These longitudinal data provide the ability to evaluate change or stability over time (Cole & Maxwell, 2003) which is a ubiquitous research question in many areas of study, such as the social sciences, biology, economics, political science, and education to name a few (Askildsen, Jirjahn & Smith, 2006; O'Donnell, 2016; Villa, 2017).

One key consideration that should not be ignored with longitudinal data is the idea that the scales or surveys used in these studies should be measuring the same thing at each time point. That is to say, the operational definition of the construct of interest should remain the same across measurement occasions. If the psychometric properties of the questionnaire differ across time or between groups, then this difference is confounded with any other observed differences due to time, intervention, or other conditions of the study. Any mean differences cannot be interpreted as true mean differences since the change may be due to changes in the measure instead of true change among individuals. This bias in measurement is what methodologists refer to as a lack of measurement invariance (LMI); although it is also known by other names such as factorial invariance, measurement equivalence, and differential item functioning (Millsap & Cham, 2012). The idea of MI may seem obvious to the seasoned researcher, however, it is often overlooked or simply assumed instead of evaluated when researchers use classical statistical techniques such as repeated measures analysis of variance (ANOVA) or other manifest variable techniques (i.e. ordinary least squares regression). However, structural equation modeling (SEM) is an analytic tool that can evaluate MI.

In this chapter I will briefly present some introductory background on CFA in the SEM framework that includes the strengths of this technique and a brief review of model fit and evaluation in order to familiarize the readers with some terms and definitions related to MI evaluation in SEM. Then, I will define and explain MI including current procedures and guidelines involved in MI evaluation. Finally, special considerations for longitudinal research will be discussed along with gaps in the literature which will lead to the purpose of this study.

Structural Equation Modeling Framework

SEM centers around a distinction between two types of variables; manifest variables and latent variables. Latent variables are unobserved constructs that researchers attempt to measure using manifest variables. Manifest variables are the typical questions researchers think of when designing a survey for a study, for example, items or questions in a survey that prompt participants to respond, complete a task, move a sliding scale, or rate how much they agree/disagree with statements (e.g., personality or behavioral measures, cognitive tasks) are then used as variables in analyses. These variables are considered observed (manifest) variables of a construct (i.e. extraversion, executive function, memory, depression). These manifest variables can then be used as indicators of latent constructs. It should be noted that SEM has also been referred to as covariance structure analysis, covariance structure modeling, and analysis of covariance structures (Kline, 2016). The use of SEM has increased in recent years due to advances in computing power, availability of SEM software programs, and the flexibility SEM provides compared to other common statistical analyses such as ANOVA and regression (Hox & Bechger, 1998). Some advantages of SEM include accounting for measurement error, fitting/analyzing complex research questions and hypotheses in a single model, and the ability to assess MI (Hoyle, 2012; Kline, 2016; Little, 2013). This paper will focus on the last benefit

mentioned, SEM's ability to evaluate MI. The foundation of an SEM is a confirmatory factor analytic (CFA) measurement model. Essentially, CFA is a special case of SEM and a CFA measurement model is specified and evaluated prior to testing a hypothesized SEM. It is within the CFA measurement model process that MI can be evaluated.

Introduction to MI

Over the last decade MI has become a hot research topic in the quantitative literature. This may be due to the increased availability and usage of SEM in applied research and the fact that researchers who use questionnaires in different subgroups can evaluate MI in this framework. Although the use of SEM still requires statistical and distributional assumptions (like any statistic) to be met, it is also considered one of the most flexible techniques (Hox & Bechger, 1998; Little, 2013). Some of the flexibility lies in the fact that some assumptions can be evaluated instead of just simply assumed. For example, one of the major strengths of SEM is the ability to actually test for MI where as other analytic methods based on the General Linear Model (GLM) such as Analysis of Variance (ANOVA) or Regression, researchers must typically just assume a measure is equivalent across groups and time.

Many research questions involve comparing groups, evaluating group differences, and/or observing behavior across repeated measurements. For example, researchers may be interested in reaction time on a cognitive or physical task, physical attributes, personality, memory, learning, behavior, health habits, development, etc. Researchers are also usually interested in comparing groups of people and/or comparing people across time. In order to conclude differences are due to group dynamics or time and not changes in the measure, we must establish that the measurement properties do not vary across groups or time. This is the goal of assessing MI; examining MI allows researchers to evaluate whether or not the operational definition of the

measure(s) being used differs or changes between groups and/or time (Kline, 2016). Doing so will allow researchers to answer the questions: Are the groups different or is it the measure? Do people change across time or does the meaning of the measure change across time? SEM allows researchers to evaluate MI directly in the CFA framework. If MI is not directly assessed or is not found tenable, then the inferences drawn about differences between groups and/or across time are likely to be obscured by differences in the measure(s) (Horn & McArdle, 1992).

Steps of Evaluating MI in CFA

The most common approach for evaluating MI in the CFA framework requires three or four steps: configural, weak, strong, and strict (Dimitrov, 2010). Each of these steps add constraints to the model that become progressively restrictive. The models are evaluated along the way in order to assess if a certain set of restrictions significantly worsens model fit. Although the literature review on MI by Vandenberg and Lance (2000) details that over the last few decades various combinations of these steps have been used, this paper will focus on the most commonly used procedures currently employed in the social sciences (Schmitt & Kuljanin 2008; Kline, 2016; Little, 2013; Dimitrov, 2010). Additionally, the procedures discussed have been developed using cross-section or multiple group study designs, but the same procedures are applied to longitudinal data (Millsap & Cham, 2012).

Configural Invariance. The first step is by testing what is called the configural invariance model (Dimitrov, 2010; Meredith 1993; Millsap & Olivera-Aguilar, 2012). The only constraints placed on the configural model are what are needed for scale setting and model identification purposes. In order for configural invariance to “pass” the model must have good overall model fit and the pattern of free and fixed factor loadings should be the same across

groups/time (Horn & McArdle, 1992). If the model fits well and if the pattern of fixed and free parameters is equivalent over time and/or between groups, then there is evidence of configural invariance. Configural invariance is represented by the following definitional equation from Horn and McArdle (1992).

$$Y_0 = T_0 + \Lambda_0\eta_0 + \Theta_0$$

Where Y are the scores on the items and subscript o refers to the occasion of measurement and can be denoted by a g when referring to multiple groups. When present, these subscripts indicate that the estimates for those parameters are allowed to freely vary across time or groups. The T represent item means vector, Λ are the lambda loadings matrix for the latent construct η and Θ represents the residual variance / covariance matrix of the items.

Evaluating Model Fit. How do we determine if a CFA model fits well? Many researchers agree that there is no single test of model fit that will inform a researcher that a SEM or CFA model fits well (Fan & Sivo, 2011; MacCallum, 2003; Millsap & Cham, 2012). They suggest that multiple fit indices should be used collectively in order to gather evidence of how a particular model fits the observed data (Sarlis, Satorra, & van der Veld, 2009). In addition to these fit indices, researchers should also evaluate the magnitude and direction of parameter estimates, and model residuals (West, Taylor, & Wu, 2012). Parameter standard errors to determine the plausibility of the estimates and model convergence criteria should also not be ignored (Little, 2013).

At their core, each of the model fit indices attempt to evaluate the difference between the observed covariance matrix S (from the data collected) and the model implied covariance matrix $\hat{\Sigma}$ (model derived based on the research hypotheses a researcher is testing). There are many model fit indices; however, here the focus is on the handful of model fit indices that are currently available as output from SEM software programs, are commonly reported for CFA models, and that have been utilized in evaluating MI (Hox & Bechger, 1998; Kim & Willson, 2014). The model chi-square statistic (χ^2), comparative fit index (CFI; Bentler, 1990), and the root mean square error of approximation (RMSEA; Steiger, 1990) will be discussed. These fit indices are also the most commonly used and reported in the applied literature in regard to CFA models in SEM (West, Taylor, & Wu, 2012). Gamma hat (Maiti & Mukherjee, 1990) will also be incorporated due to its ability to account for the number of items in a model.

Model Chi-Square. The χ^2 in SEM is a statistical fit index that is commonly used by researchers as a test of model fit. The χ^2 statistic is a test of “exact” fit. In SEM, the model χ^2 represents the plausibility of the null hypothesis that the observed variance/covariance matrix is equal (statistically equivalent) to the model implied variance/covariance matrix ($\Sigma = \Sigma(\theta)$). The research hypothesis, then, is the opposite; the variance/covariance matrices are different. In SEM, the goal is to fail to reject the null hypothesis for the model chi-square statistic. Researchers want the observed information (data) to be approximately equal to the model implied information (hypothesized structure). The model implied is the model that is fit to the data based on research hypotheses (our hypothesized model). However, many researchers argue that the χ^2 statistic is sensitive to sample size (Brown, 2015; Kline, 2016; Little, 2013; Dimitrov, 2010; West, Taylor, & Wu, 2012) and minor deviations from exact fit (King-Kallimanis, Oort, & Garst, 2010). Additionally, SEM is a large sample technique so most models require larger

samples sizes compared to manifest variable frameworks (e.g. ANOVA, OLS regression). Therefore, more often than not in practice, the χ^2 test of statistical fit is significant, which is not typically desired in this framework. Larger sample sizes have more power and miniscule differences between the variance/covariance matrices may be “statistically” significant which can lead to many reasonable models being rejected (West et al., 2012; Little, 2013). This is similar to what happens with simple independent samples t-tests with large sample sizes. The χ^2 is conceptually appealing, but overly influenced by sample size. Although, the χ^2 alone should not be used to evaluate model fit, it is the basis of other fit indices in SEM and can be used to test for nested model comparisons.

Alternative Fit Indices. Due to the perceived sensitivity to the chi-square test of exact fit and the argument that exact fit could be unrealistic in the SEM or modeling framework (“all models are wrong, some are useful” Box, 1979, pp. 202), some alternative fit indices (AFIs) have been created. One commonly reported AFI is the comparative fit index (CFI) The CFI is considered a relative fit indice and it compares a predicted (tested) model with a baseline (null or worst fitting) model in order to evaluate how improved the predicted model is compared to the null. The CFI (Bentler, 1990) provides a ratio of misfit from the tested model (χ_T^2) compared to a baseline model (χ_B^2):

$$CFI = 1 - \frac{\max(\chi_T^2 - df_T, 0)}{\max(\chi_T^2 - df_T, \chi_B^2 - df_B, 0)}$$

where χ_T^2 is the tested model’s χ^2 , df_T is the degrees of freedom for the tested model χ^2 , χ_B^2 and the df_B are χ^2 and degrees of freedom for the baseline (typically the default independence or null) model. The max operator signifies that the maximum of the values separated by commas

will be used. Values of 1.0 are considered perfect fit while less than .90 are considered poor fitting models.

Another AFI commonly reported in the literature is the root mean squared error of approximation (RMSEA; Steiger, 1989). The RMSEA is a measure of absolute fit. Absolute fit indices compare the predicted model with the perfect fitting model, without reference to a poor fitting model. Values closer to zero are interpreted as better fitting models; an RMSEA of zero would indicate perfect fit. Given this interpretation, the RMSEA can be considered a badness-of-fit index (West, Taylor & Wu, 2012). RMSEA values between .05 - .08 are considered acceptable while higher values may be evidence of a poor fitting model. Most SEM software also provide a Confidence Interval (CI) for the RMSEA; this is unique in that it is the only commonly used fit index with a CI. Kenny, Kaniskan, and McCoach (2014) argue that RMSEA cannot be trusted (may be biased) in models with small degrees of freedom. Fan & Sivo (2007) also found that the RMSEA over performs in larger models (many indicators) indicating good fit for a poor fitting model. The RMSEA is calculated as follows:

$$RMSEA = \sqrt{[(\chi^2_T - df_T)/(N - 1)] / (df_T/g)}$$

Where N is the sample size and g represents the number of groups.

Another global fit index that might be useful in evaluating fit and MI is Gamma Hat (Maiti and Mukherjee, 1990; West, Taylor, & Wu, 2012). Values closer to 1.0 indicate better fit (Fan & Sivo, 2007). Although Cheung and Rensvold (2002) included Gamma hat in their MI study and recommended a difference of less than .001; it has not been reported in studies evaluating MI. One reason for its exclusion could be due to SEM software defaults; Gamma Hat

is not one of the fit indices automatically provided in the output but may be calculated using the formula:

$$\text{gammaHat} = \frac{p}{p + 2 \left(\frac{\chi^2_T - df_T}{N} \right)}$$

where p is the number of observed variables in the model.

All of these fit indices should be evaluated in order to determine if the configural model fits acceptably well based on the guidelines mentioned. A word of caution should be noted here; these cutoffs should be interpreted as rough guidelines as they have also been criticized (McDonald & Ho, 2002). Researchers should also make informed decisions based on previous literature in their particular field of study. Methodologists have created and/or adjusted many more fit indices that can be used to evaluate model fit (e.g. Tucker-Lewis Index, Akaike information criterion, and Bayesian information criterion). However, and as previously mentioned, this paper will focus on the most commonly used and reported fit indices for nested model comparisons that also have a history in MI evaluations.

Weak Invariance model. Weak invariance is the next step of MI (Widaman & Reise, 1997). Also referred to as metric invariance (Horn & McArdle, 1992) or pattern invariance (Millsap, 2011), weak MI is evident if the corresponding factor loadings (λ) are equivalent across time/groups. The weak invariance model with the addition of these constraints, imposed by the researcher, is nested within the configural model. As such, nested model comparisons can be used to evaluate if the constraints are tenable. Two models are said to be nested if one model can be transformed into the second model by imposing constraints on the parameters (Chou & Bentler, 2002; Chou & Huh, 2012). In order to statistically test each model, fit indices and model comparisons are evaluated. Nested models can be compared using the model chi-square

difference statistic ($\Delta\chi^2$ and Δdf). However, previous research suggests using the CFI and RMSEA as well (Cheung & Rensvold, 2002; Little, 2013). There is evidence to support MI if the change in CFI (ΔCFI) does not decrease more than .01 based on a simulation study by Cheung and Rensvold (2002). Another study by Meade et al. (2008) recommends a more conservative ΔCFI of less than .002, but this recommendation is not as commonly used in the applied literature. Little (2013) suggests that the CI of the RMSEA for the nested models can also be used to evaluate MI. If the RMSEA indices for the nested model fall within the CIs, then there is evidence of MI. The definitional equation for weak invariance is below. You will note that the only minor change to the formula is the absence the subscript after the lambda (Λ) indicating that each occasion has been constrained to equality across time or group for those parameters.

$$Y_0 = T_0 + \Lambda\eta_0 + \Theta_0$$

Strong Invariance. If weak invariance holds, the next step in evaluating MI is the strong (Meredith, 1993) or scalar (Steenkamp & Baumgartner, 1998) invariance model. This model retains all previous constraints and adds equality constraints of the indicator intercepts, Note that the formula below no longer has the subscript for the Tau matrix (T). The weak and strong models are also nested and can be tested using the nested model comparisons mentioned above. Strong MI is required if researchers are interested in mean differences / changes between groups or across time.

$$Y_0 = T + \Lambda\eta_0 + \Theta_0$$

Strict Invariance. Finally, the last test of MI is strict invariance. Strict invariance is met if the corresponding indicator residuals are equivalent (Meredith, 1993), see formula below. There is some disagreement in the literature as to whether or not to impose a strict level of invariance (Millsap & Olivera-Aguilar, 2012). DeShon (2004) suggests enforcing this level of MI since the item specific variance of each indicator should be equivalent across groups and could complicate mean differences if not constrained. However, other researchers do not recommend such conservative rules at this level because it is a difficult and unrealistic criterion to meet with real data (Little, Card, Slegers & Ledford, 2007; Little, 2013; see also Brown, 2015; Chan, 1998), deemed less important and/or least interesting in most research (Widaman & Reise, 1997; Dimitrov, 2010), and overly restrictive (Bentler, 2004; Byrne, 1988). Keep in mind that the residual term in SEM contains both item specific variance in addition to random noise/error variance (Bollen, 1989; Bollen & Hoyle, 2012). In most (if not all) research it is unreasonable to expect that random error variance would be equivalent between groups and/or across time since it is situation specific variance and does not belong to the measurement instrument.

$$Y_0 = T + \Lambda\eta_0 + \Theta$$

Once MI (whether it be across time and/or group) is established, then researchers can start to test hypotheses about latent means, variances, and relationships (covariances/regressions) of the latent constructs. Kline (2016) considers this level of invariance testing Structural Invariance (SI; see also Little, 2013 or Dimitrov; 2010). However, MI must be evaluated before moving onto SI. If measurement invariance is not tenable, then differences found in the latent

space may be compromised by differences in the measurement instrument instead of the differences actually occurring due to time. (Millsap & Olivera-Aguilar, 2012).

Special Considerations in Longitudinal CFA and MI Evaluation

Longitudinal studies, also called repeated measurements, enable us to answer important questions about change over time. However, researchers generally assume the measures they are using at each of these repeated assessments are measuring the same construct each time.

Anytime we want to examine or compare a measure across groups or across time points we are usually assuming that the measure is invariant across time and/or groups. This means that the measure is actually measuring the exact same construct across time (Millsap, 2011). This assumption is made anytime we do a repeated measure or between groups ANOVA. The ANOVA framework does not actually allow us to test this assumption nor do the regression frameworks. Whereas, this assumption can be tested (and justifiably relaxed) in SEM.

Although some applied research evaluated longitudinal MI using guidelines provided by multiple sample MI simulation studies (Barclay-Goddard et al., 2009; King-Kallimanis et al., 2011; Oort, Visser, & Sprangers, 2005;), others fail to evaluate longitudinal MI at all (Brumley, Brumley, & Jaffee, 2019; Burkholder & Harlow, 2003; Spinath & Spinath, 2005; Thompson, Sims, Kingree, & Windle, 2008; Vautier, 2004) before testing hypothesized structural relations. Even though numerous studies have examined the negative impact of failing MI on parameter estimates, model fit, and study conclusions in longitudinal SEM (e.g., Bishop, Geiser, & Cole, 2015; Ferrer et al., 2008; Leite, 2007; Millsap & Cham, 2012; Olivera-Aguilar, 2013), optimal strategies for testing longitudinal MI are unclear and further research is imperative (Millsap & Cham, 2012).

Longitudinal CFA and SEM models add complexity. One latent construct measured at one time point is a very simple model, even if it is measured between two-groups. Figure 1 provides an example of a single latent construct (η) measured by four variables (V , also called items or indicators in the literature). The other aspects of Figure 1 are common diagramming conventions used to represent a construct in CFA or SEM. The single headed arrows represent regression type parameters and here they represent the factor loadings (λ) for each indicator. Double headed arrows represent the variance of an indicator (θ) or construct (Ψ) with itself or covariances / correlations.

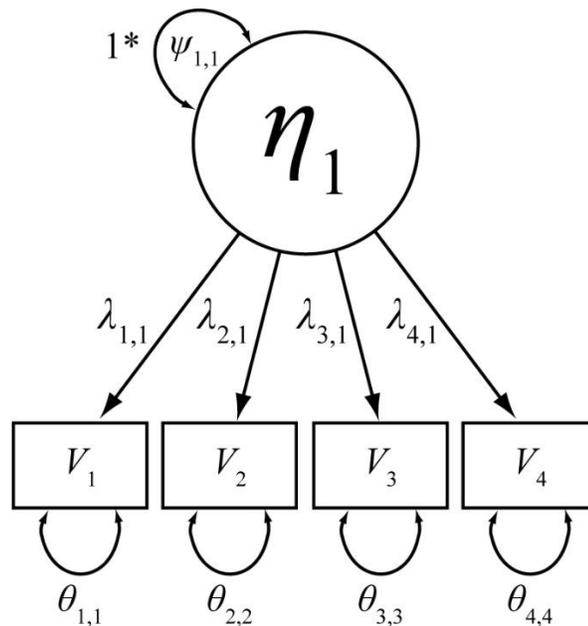


Figure 1. Single Latent Construct Example

Adding multiple occasions to this simple single construct increases the complexity. Not only is the latent construct now allowed to correlate with itself across time, but each indicator of the latent construct is allowed to covary with itself across time as well (see Figure Two for example). If we measure across four occasions, that adds six newly estimated parameters. That

means if we are measuring self-determination among adolescents using four questions across four measurement occasions, we estimate an additional 24 (4 items x 6 covariances) parameters.

These parameters do not exist if we were examining multiple groups at a single time point.

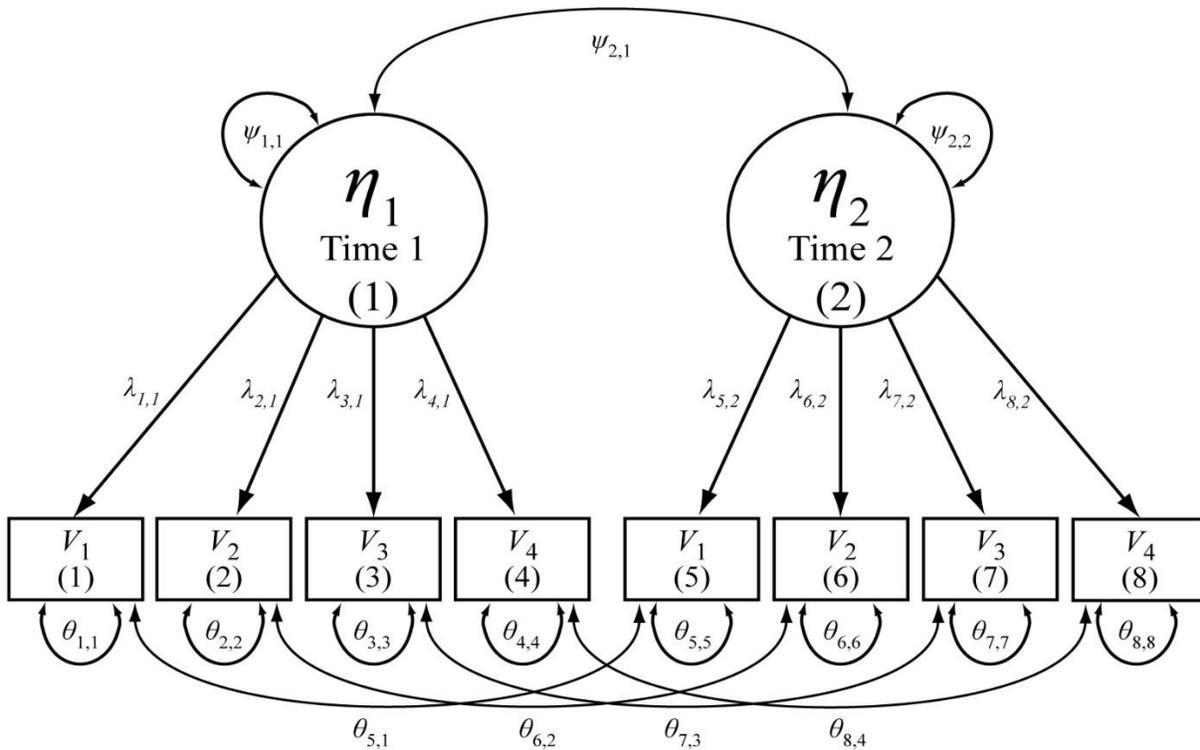


Figure 2. Two Repeated Measurement Occasions

Another consideration in longitudinal CFAs relates to the null model. In order to evaluate longitudinal MI, we must start with an appropriate null model. Widaman and Thompson (2003) point out that the default null model (also referred to as the independence model) in SEM software is not correct. The standard ‘null’ model assumes that all covariances are zero and only variances are estimated. In longitudinal and multiple group research, a more appropriate ‘null’ model is to assume the variances of each corresponding indicator are equal at each time point

and across groups in addition to their means (intercepts) also being equal at each time point and between groups. Additionally, any covariances across time should be equated in the null model (Widaman & Thompson, 2003).

In longitudinal SEM, one of the first models to be fit (after the appropriate null model) should be a longitudinal CFA that includes each occasion to evaluate the longitudinal model (Little, 2013). This model is equivalent to a longitudinal configural CFA model. Longitudinal MI is then tested using the same techniques and steps described previously (by testing configural, weak, and strong invariance, respectively).

Current Limitations

Many researchers have studied multiple group invariance (for a review see French & Finch, 2006; Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008) and although the nested models can be compared using $\Delta\chi^2$, many researchers have noted that the chi-square difference test is also sensitive to sample size (Brown, 2015; Kline, 2016; Little, 2013; Dimitrov, 2010; West, Taylor, & Wu, 2012) especially when testing for MI (Meade & Bauer, 2007; Little, 2013) and therefore also rely on changes in AFIs to evaluate MI (Putnick & Bornstein, 2016).

Reporting multiple pieces of evidence, such as $\Delta\chi^2$, ΔCFI , and the RMSEA point estimate and CI, when evaluating MI is becoming a more standard practice (Dimitrov, 2010). A recent review of 126 applied research articles published between May 2013 and April 2014 that evaluated MI was conducted by Putnick and Bornstein (2016). Out of the 126 articles reviewed 16.7% reported $\Delta\chi^2$ only, 34.1% reported only $\Delta AFIs$, and 45.9% reported both when evaluating MI. The remaining 3.3% only reported no change in model fit.

Current commonly used recommendations and guidelines for testing MI are based on multiple group studies and simulations. For example, Cheung and Rensvold (2002) evaluated 20

different measures of fit in order to determine if change in fit could be used for evaluating invariance in a multiple group framework. Out of the 20 fit indices examined, two that were recommended are the change in the CFI and change in gamma hat when comparing models. The CFI is provided in most modeling software, so it has become a popular way to evaluate MI and based on Cheung and Rensvold's recommendations a decrease in the CFI greater than .01 is an indication of failure of invariance (items are variant). These indices were evaluated using only two-groups but have been used and recommended with longitudinal MI testing as well (Little, 2013). Little also recommends reporting the RMSEA point estimate and CI and provides the guideline that the RMSEA value should fall within the CI for each nested model. Although, no empirical support for this recommendation has been reported in the literature.

This study builds upon the previous MI research in the following ways. (1) While previous research has shown the negative outcomes of failing MI in longitudinal data (e.g., Bishop, Geiser, & Cole, 2015; Ferrer et al., 2008; Leite, 2007; Millsap & Cham, 2012; Olivera-Aguilar, 2013), clear guidance for longitudinal evaluation of MI is lacking (Millsap & Cham, 2012). This study evaluates MI detection for a latent construct across 2, 4, 6, and 8 repeated measurement occasions. (2) While current guidelines exist for testing MI (Cheung & Rensvold, 2002; Meade, Johnson & Braddy, 2008; Chen, 2007), they were developed in the context of multiple-group models even though they have been cited numerous in longitudinal research for assessing MI across multiple time points (e.g., Motl & DiStefano, 2002; Obradović, Pardini, Long, & Loeber, 2007). Multiple-group models differ from longitudinal models in a number of key aspects, which may influence the determination of MI. Therefore, this study examines if the current cut-offs are sufficient in the longitudinal case. (3) Widaman and Thompson (2003) proposed that the null model used in the calculation of incremental fit indices (e.g., CFI, Bentler,

1990) that is generated by default in software packages is flawed. Evaluation of these cutoffs using the appropriate longitudinal null model proposed by Widaman and Thompson (2003) has not been considered. The CFI values using both the default null (independence) model and the appropriate null are model are evaluated in order to consider the significance of the Widaman and Thompson (2003) article for MI.

Simulation Studies

Simulation studies are computer generated data experiments. Instead of collecting data from participants, data are generated based on specifications laid out by the researcher. Simulating (generating) data using specific model and population parameters allows researchers to study statistical procedures (Burton et al., 2006). Specifically, researchers are able to determine how well a statistical procedure performs under different circumstances experienced in applied work. These studies are experimentally designed, and researchers can systematically manipulate different variables or parameters of interest in order to evaluate how well a procedure works under different conditions (e.g. assumption violations, failing to meet a criterion, evaluating different estimators). Data problems can be simulated and tested across hundreds (or thousands) of data sets in order to see how trustworthy or biased the results may be since with simulated data the true values are known. For an introduction to simulation studies interested readers are directed to texts such as Monte Carlo Simulation and Resampling Methods for Social Science by Carsey & Harden (2013). Simulations are also used to determine sufficient sample sizes before researchers collect data from participants in order to help determine the number of participants needed to detect effects (power analyses). Conducting a simulation study allows evaluation of the current guidelines as applied in the longitudinal case in order to fill the gap in the literature in regards to MI evaluation in longitudinal CFA.

Purpose of the Current Study

The goal of this study is to begin providing researchers with guidance on how the alternative fit indices and difference in chi-square perform when used to evaluate MI in the longitudinal case.

Specifically:

1. How does the $\Delta\chi^2$ perform in evaluating longitudinal MI?
2. How does the change in CFI guideline hold up in longitudinal MI?
3. Does the specification of the appropriate null model affect ΔCFI guidelines for evaluating longitudinal MI?
4. Is the CI for RMSEA able to detect MI under the conditions in the current study?
5. Will there be any differences in MI detection based on the number of repeated occasions of measurement?

Chapter 3: Method

A Monte Carlo simulation study was used to answer the above research questions. This section details how the data were generated for the population models, the conditions that were manipulated, and which parameters were fixed (non-varying). The outcomes of interest, Type I error rates and power of model fit indices, are also described.

Data Generations

Invariant models. Invariant data generating models included one latent construct measured by seven indicator variables over repeated measurement occasions. The first models for data generation did not include LMI, meaning that the factor loadings and item intercepts of each indicator were not variant. The no LMI condition models are used to examine how often LMI is detected when there is none (Type I error rates). Other fixed population values for each model followed common specification from previous research using CFAs (Cheung & Rensvold, 2002; French & Finch, 2006; Kim & Wilson, 2014). Factor loadings (λ) were set at .7, so the latent construct accounts for approximately 49% of the variance in each item, (λ^2). Residuals (θ) were fixed at 0.51 based on the fixed factor loadings ($1 - \lambda^2$) and item intercepts were fixed at zero. Latent construct relations and indicator residuals over time followed an auto regressive (AR1) simplex structure in which latent correlations at adjacent time points started at .80 while indicator residuals started at .30 and then decreased over time (see Figure 3 for the population model with two-time points). These longitudinally measurement invariant data generation models included two, four, six, and eight-time points across two sample sizes ($N = 400, 800$) for a total of 8 non-variant conditions.

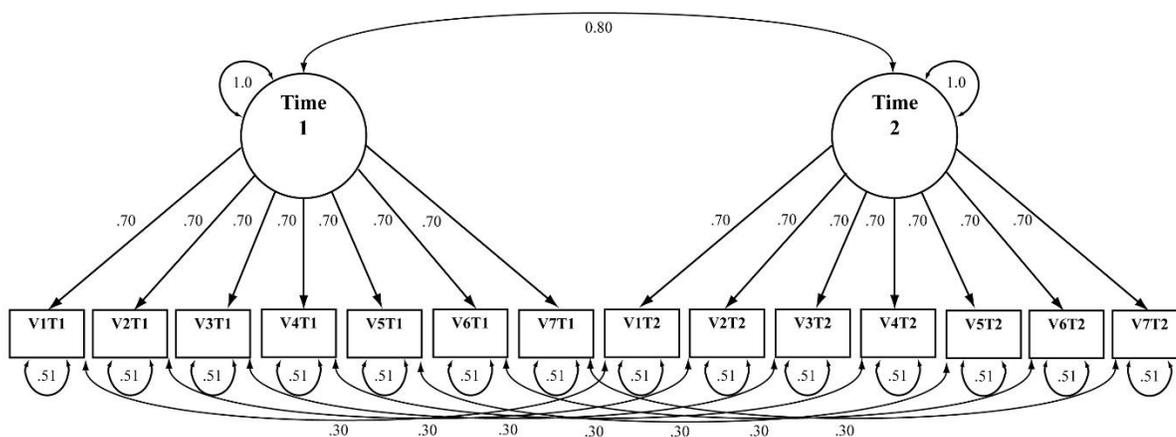


Figure 3. Invariant Model with Two Occasions

Variation model simulation conditions. The conditions (32) were manipulated in order to evaluate the performance of $\Delta\chi^2$ and ΔAFI s' ability to accurately detect LMI in longitudinal CFA (power) are described below. Table 1 summarizes the manipulated conditions, variable names used, and the levels.

Table 1.

Manipulated Variables in Monte Carlo Design

Variable	Description	Levels
Type	Type of LMI	factor loadings or item intercepts
Magnitude	Magnitude of LMI	small or large
Time	Number of time points	2, 4, 6, or 8
N	Sample size	400 or 800

Type of LMI. Conditions to test for the type of invariance included factor loadings or item intercepts. LMI in the factor loadings represents a lack of weak invariance. Weak invariance is

the minimal requirement needed in CFA that allows researchers to examine latent variances and covariances across groups or time. LMI in the item intercepts represents a lack of strong invariance and strong invariance is required in order to examine latent means.

Magnitude of LMI. Across the population models, the magnitude of non-invariance varied between small amounts of non-invariance and large amounts of non-invariance (2 levels). The magnitude for a small amount of non-invariance was 0.15 decrease in factor loadings. Figure 4 provides the diagram of the population model for two occasions; note that factor loadings for variables five, six, and seven for time two are .55. The magnitude for a small amount of non-invariance in the item intercepts was 0.25 and can be seen in Figure 5. The magnitude for large amounts of non-invariance was 0.3 for factor loadings and 0.5 for item intercepts. These magnitudes are standardized and in line with Kim and Wilson's (2014) study on multiple group invariance. These magnitudes also relate to small and moderate effect sizes (Cohen, 1992; Ferguson, 2009). Previous simulation research has failed to examine methods for evaluating longitudinal MI independent of multiple groups, so this study used the current research values as a starting point for investigation.

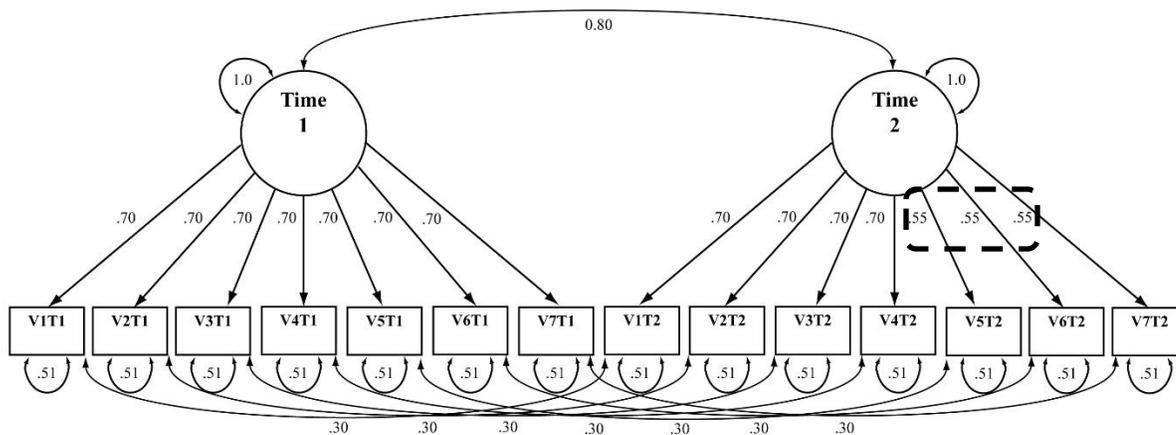


Figure 4. Small Loading LMI

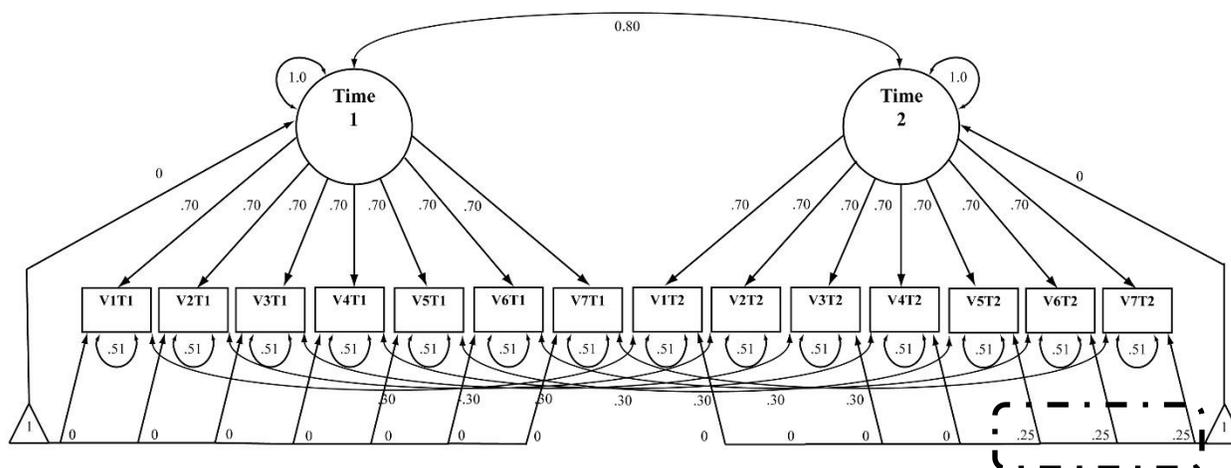


Figure 5. Small Intercept LMI

Time points. The number of time point conditions ranged from two to eight measurement occasions by a factor of 2 (2,4,6,8). LMI occurred on the last three indicators (variables/items five, six, and seven) at the second, third, fourth, and fifth time point, respectively, and remained invariant from initial values until the final time point in the model. LMI of this form can be seen in applied research that depicts evidence of developmental or response shift (Ahmed et al., 2005; Fokkema, Smits, Kelderman, & Cuijpers, 2013; King-Kallimanis, Oort, & Garst, 2010; Schwartz & Sprangers, 1999; Schwartz, Sprangers, Carey, & Reed, 2004) in which once an item is found to be invariant, it remains invariant at future measurement occasions.

As previously mentioned, correlations across time followed an AR1 simplex structure. Specifically, latent correlations started at .80 for adjacent timepoints, and decreased over time so that time 1 correlated at .64 ($.80 * .80$), .51, .41, .33, .21, and .17 with occasions three through eight, respectively. A similar pattern was set for the item residuals and correlations across time started at .30 for adjacent timepoints.

Sample size. The sample size condition has 2 levels ($N = 400$ and $N = 800$) in order to have sufficient sample sizes for the models in this study (Maxwell, 2004; Singer & Willett, 2003; Wall, Guo, & Amemiya, 2012).

Implementation of data generation. Complete multivariate normal data were generated in R (R Core Team, 2018) using the *simsem* package (Pornprasertmanit, Miller, & Schoemann, 2016). Symmetric, equally spaced, categories in line with Rhemtulla et al., (2012) were created to mimic 7-point Likert type response items (thresholds = -2.5 -1.5 -0.5 0.5 1.5 2.5). One thousand replications were generated in each condition ($R = 40,000$).

Data Analysis Models

Alternative null model. In addition to using the software default independence model (i.e., constrain all covariances to zero, but freely estimate all variances and means), to calculate the CFI, the alternative longitudinal null model in which variances and means are also constrained to equality across time (Widaman & Thompson, 2003) was fit to each data generation model in order to use the alternative baseline chi-square to calculate the CFI_a^1 for each analysis model.

Measurement invariance analysis models. Analysis models included the CFA models described previously that are used to evaluate MI (configural, weak, and strong invariant models). The configural invariance analysis model for each condition were estimated to have the same pattern of free and fixed parameters. The variance of the latent construct was fixed to 1.0 and the latent mean was fixed at zero to set the scale. All other parameters were freely estimated

¹ The subscript “a” is used to denote that the fit index was calculated using information from the alternative null model.

including latent correlations over time, factor loadings, item intercepts, and item residuals. Item residual correlations among the same items over time were also freely estimated.

The weak invariance analysis models followed the same structure except the factor loading for each item was constrained with the same item over time. Additionally, once these constraints are added to test for MI, the restrictions to set the scale are only needed for the first timepoint and the following timepoints are freely estimated.

Strong invariant analysis models for each condition added equality constraints on the item intercepts over time and latent means are freely estimated. All other parameter estimation matched the previous (weak) model specifications. The strict invariant model was also estimated for the invariant conditions. The strict model imposes additional equality constraints on the residuals over time.

Model fit information (the scaled chi-square, CFI, CFI_a, RMSEA with CI, and gamma hat) was saved from all replicated analyses in order to calculate the model difference tests. These model fit indices were used to calculate Δ chi-square and Δ AFIs by comparing the configural versus weak models and the weak versus strong models and the strong versus strict models in order to evaluate invariance detection.

Implementation in R. All conditions analyzed with the four models were implemented in lavaan (version 0.5; Rosseel, 2012). The scales of the latent factors were identified by using the fixed factor method of scaling (latent factor variance fixed to one and the latent mean to zero). A robust ML estimator (Bentler & Satorra, 2010) was used to account for the nonnormality induced by the Likert-type data. Due to the use of the robust estimator, the scaled chi-square was used in model evaluations and comparisons.

Outcomes

Type I error and power are the outcomes of interest in most simulation studies. In this study, Type I error rates (false positives) were evaluated for the $\Delta\chi^2$ and $\Delta AFIs$ in line with Cheung and Rensvold (2002). In order to determine Type I error, data were simulated that follows the null hypothesis of invariance. In this study, Type I error is an indication of a false identification of non-invariance (items are found to be variant when they are not). Power to detect longitudinal MI with the $\Delta\chi^2$ and $\Delta AFIs$ was evaluated in conditions where LMI was present. Power for the $\Delta AFIs$ was computed by taking the number of replications in which LMI was detected using the current guidelines previously discussed divided by the number of replications for each condition (Bandalos & Gagné, 2012).

Chapter 4: Results

The outcomes of interest in this study are the type I error rates when MI holds (no variant items) and power to detect LMI when present by examining the difference in the chi-square statistic and the AFI's based on guidelines provided in the literature. The proportion of the 1,000 replications in which a given test returned a false rejection represents the Type I error rate per condition. In conditions where MI does not hold, this value represents the Power of each test to detect LMI. All replications converged in this simulation study.

Type I Error Rates of the scaled $\Delta\chi^2$ and ΔAFI 's

The scaled chi-squared difference test has Type I error rates of zero for the configural versus weak model comparisons. Type error increase to expected rates ranging between .05 to .07 when comparing weak versus strong and strong versus strict models (see Table 2.). Sample size did not seem to influence the scaled chi-square difference test in this study. Meade et al., (2008) suggestion of the CFI change of less than .002 guideline had Type I error rates that ranged from .02 to .07 for sample size of 400 and were lower for the larger sample size (N=800; range 0 - .02). The change in CFI (.002) using the alternative null model proposed by Widaman and Thompson were almost identical to the default independence null provided by the software. The error rates were the highest for the change in Gamma hat and ranged from .12 to .20 in the smaller sample condition (N= 400) but were lower (range .01 to .09) and closer to acceptable Type I error ranges in the larger sample size conditions (N = 800). The lowest error rates for the change in Gamma hat were found in the eight occasion conditions.

Table 2.

Type I Error Rates

		N = 400				N = 800			
		T = 2	T = 4	T = 6	T = 8	T = 2	T = 4	T = 6	T = 8
Chi-square	Configural vs Weak	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Weak vs Strong	0.06	0.06	0.07	0.06	0.04	0.05	0.05	0.07
	Strong vs Strict	0.05	0.05	0.05	0.05	0.04	0.06	0.06	0.06
CFI .002 change	Configural vs Weak	0.05	0.04	0.02	0.03	0.00	0.00	0.00	0.00
	Weak vs Strong	0.06	0.04	0.04	0.03	0.01	0.00	0.00	0.00
	Strong vs Strict	0.07	0.05	0.03	0.03	0.02	0.00	0.00	0.00
CFIa .002 change	Configural vs Weak	0.05	0.04	0.02	0.03	0.02	0.00	0.00	0.00
	Weak vs Strong	0.06	0.04	0.04	0.03	0.01	0.00	0.00	0.00
	Strong vs Strict	0.07	0.05	0.03	0.03	0.02	0.00	0.00	0.00
Gamma hat .001 change	Configural vs Weak	0.18	0.17	0.14	0.12	0.09	0.05	0.02	0.01
	Weak vs Strong	0.19	0.16	0.16	0.14	0.06	0.05	0.03	0.02
	Strong vs Strict	0.20	0.20	0.16	0.14	0.07	0.07	0.04	0.03

Note: Scaled chi-square used. N = sample size. T = number of repeated occasions

The model level Type I error rate was zero for both the RMSEA C.I. assessment and the CFI change of less than .01 between all nested models, and across all timepoints (2, 4, 6, 8) and sample sizes (400, 800). The alternative null model proposed by Widaman and Thompson (2003) did not influence the Type I error rate for the change in CFI.

Power of the scaled $\Delta\chi^2$ and ΔAFI 's

The large LMI conditions for the loadings and intercepts were large enough that all evaluation methods had sufficient power (ranging from 95 – 100%) to detect non-invariance

regardless of number occasions or sample size. Additionally, the alternative null model did not influence power to detect change in CFI.

Small LMI in Factor Loadings. The small LMI condition was a decrease in factor loadings from .70 to .55 (a change of .15). The configural versus weak invariance model comparisons were tested for all conditions. The power for the difference in the scaled chi-square ranged between .78 and .91 for the sample size of 400. The range was higher (.99 - 1.00) for the larger sample condition (N=800). The change in CFI using .01 cutoff for evaluation was not able to detect the LMI in any of the occasion conditions nor sample sizes and reached zero after by the sixth time point condition. However, the Mead et al., (2008) cutoff for the CFI change of .002 did have sufficient power across time points and sample sizes (see Figure 6). Gamma hat was also sufficiently powered to detect LMI in all conditions with power ranging from 93 to 100 percent. Power was not sufficient using the RMSEA's C.I. for evaluation and although power increased in the larger sample condition, it was only 53% for two occasions and decreased down to 10% in the eight-occasion condition (see Table 3).

Large LMI in Factor Loadings. The large LMI condition was a decrease in factor loadings from .70 to .40 (a change of .30). Power for the difference in the scaled chi-square, the drop of .002 in CFI, and the change in gamma hat were all one. Power was lowest for the change of .01 in CFI guideline and the RMSEA CI recommendation and seems to decrease slightly over time when N = 400, but power still ranged from .93 to .98 for these AFIs. Almost all tests had perfect power in when N = 800 with the slight exception that the change in CFI of .01 had power of .98 for the 6th and 8th occasion of measurement condition (see Table 3).

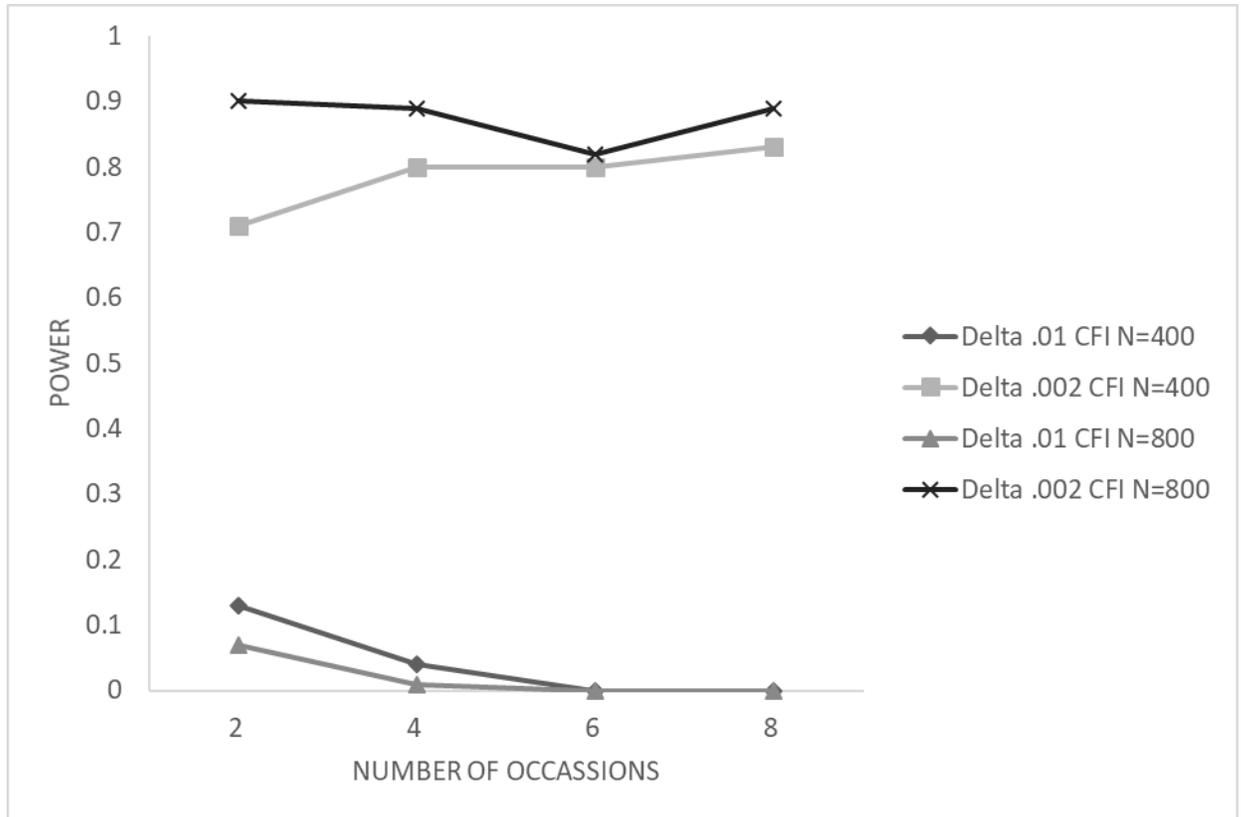


Figure 6. Change in CFI for Loadings

Table 3.

		N = 400				N = 800			
		T = 2	T = 4	T = 6	T = 8	T = 2	T = 4	T = 6	T = 8
Small loading condition (0.15)	Delta Chi SQ	0.78	0.91	0.87	0.90	0.99	0.99	1.00	1.00
	Delta .01 CFI	0.13	0.04	0.00	0.00	0.07	0.01	0.00	0.00
	Delta .002 CFI	0.71	0.80	0.80	0.83	0.90	0.89	0.82	0.89
	Delta .001 Gamma hat	0.93	0.97	0.95	0.95	0.99	1.00	0.99	1.00
	RMSEA CI	0.09	0.03	0.00	0.00	0.53	0.49	0.17	0.10
	Delta Chi SQ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Large loading condition (0.30)	Delta .01 CFI	0.98	0.97	0.95	0.95	1.00	1.00	0.98	0.98
	Delta .002 CFI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Delta .001 Gamma	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	RMSEA CI	0.96	0.99	0.93	0.93	1.00	1.00	1.00	1.00
	Delta Chi SQ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: Scaled chi-square used. N = sample size. T = number of repeated occasions

Small LMI in Item Intercepts. The small LMI condition for the intercepts was a change of .25 in the last three indicators. The weak versus strong invariance model comparisons were tested across all conditions. The scaled chi-square difference test and the change in both Gamma hat and the CFI (.002) all had power of 100% to detect LMI (see Table 4). The RMSEA C.I. method of MI evaluation only had sufficient power to detect LMI in the smaller sample (N = 400) condition for two and four occasions of measurement. Power decreased below general guidelines (<80%) for the six and eight occasion conditions. Power was 100% for the RMSEA C.I. in the larger sample size condition (N = 800). Power for the CFI change using the .01 criteria ranged from 53 to 99% with lower power in conditions with more occasions. Although the larger sample size condition had slightly higher power compared to the smaller sample condition, power fell below 80% at the four and six occasion conditions for the smaller and larger sample size conditions, respectively (see Figure 7).

Large LMI in Item Intercepts. The large LMI condition for the intercepts was a change of .50 in the last three indicators. The weak versus strong invariance model comparisons were again tested across all conditions. Power was 100% for all fit indices evaluated under this condition regardless of number of occasions or sample size. The large LMI in the intercepts condition was not able to distinguish any differences amongst power and all indices were overpowered.

Table 4.

		N = 400				N = 800			
		T = 2	T = 4	T = 6	T = 8	T = 2	T = 4	T = 6	T = 8
Small intercept condition (0.25)	Delta Chi SQ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Delta .01 CFI	0.91	0.72	0.61	0.53	0.99	0.86	0.68	0.52
	Delta .002 CFI	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	Delta .001 Gamma hat	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	RMSEA CI	0.92	0.84	0.74	0.65	1.00	1.00	1.00	1.00
	Delta Chi SQ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Large intercept condition (0.50)	Delta .01 CFI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Delta .002 CFI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Delta .001 Gamma	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	RMSEA CI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Delta Chi SQ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: Scaled chi-square used. N = sample size. T = number of repeated occasions

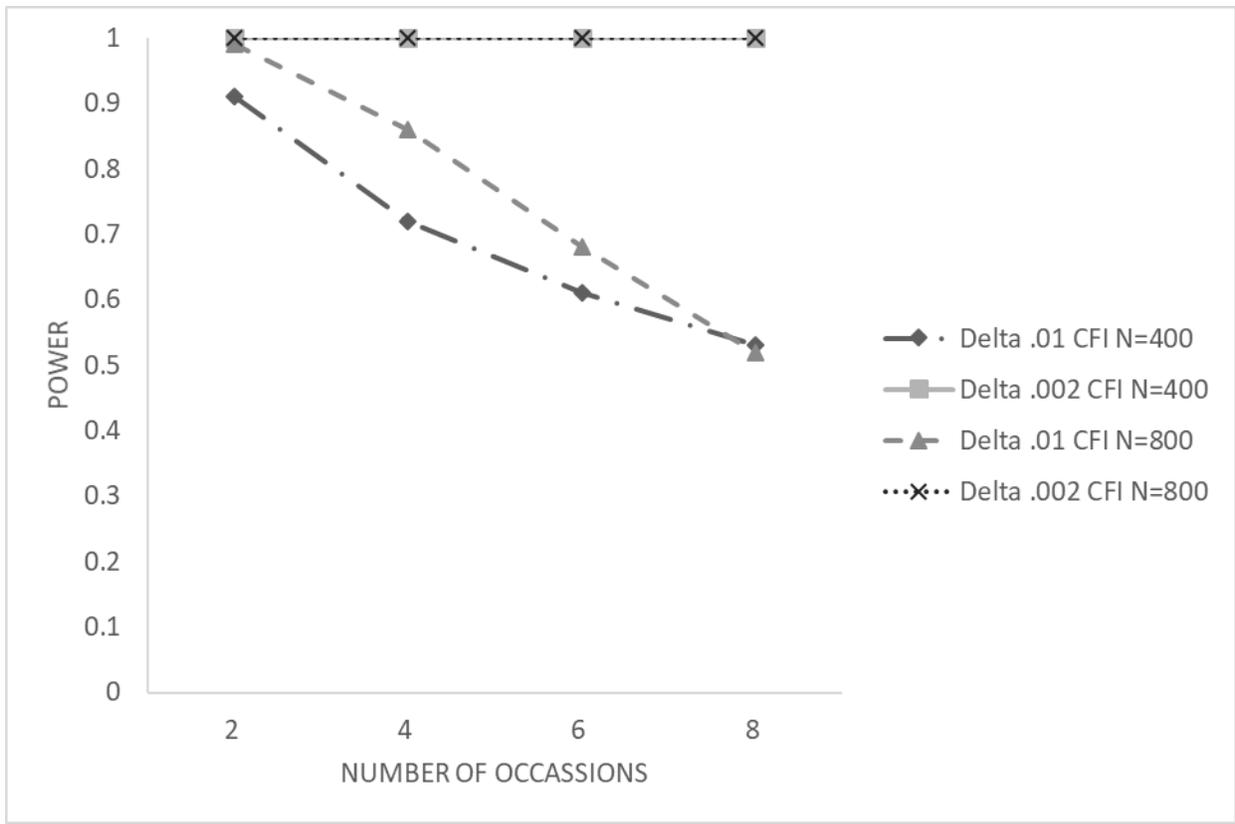


Figure 7. Small Intercept CFI Change

Chapter 5: Discussion

The goal of this simulation study was to examine the performance of what are currently the most common guidelines found in both simulation studies and applied research for evaluating MI using longitudinal CFA. There is a stark gap in the literature in this area and current suggestions stem from a small set of simulations evaluating limited cross-sectional designs. Longitudinal CFAs have some unique properties (such as, large correlations at adjacent occasions, correlated residuals for repeated items, and an exponential increase in the number of indicators in the model as number of occasions increase) compared to its cross-sectional counterparts. This chapter discusses the results based on the initial research questions followed by some ideas for future research and conclusion.

Research Questions

Performance of Chi-square difference. Type I error rates were within a range from zero to 7 percent in the invariant conditions across time, sample size, and for each MI model comparison. This is expected since there are no model misspecifications in these simulation conditions. Additionally, the scaled chi-square difference did have enough power to detect LMI in all variant conditions. Power was stronger in the variant loading condition with the larger sample size ($N=800$) and was essentially 100% even in the small intercept variant conditions. Since the chi-square is a test of exact fit, it is sensitive to minor misspecifications such as parsimony error. Parsimony error was not included in the current simulation, but is an area for future research.

Performance of the Change in AFI's. Type I error rates for all conditions were in the standard range around 5% with the exception of Gamma hat in the sample size of 400 condition.

The higher Type I error rates for gamma hat may partly explain why it is not commonly used or reported in the literature. Type I error was within range and improved at the larger sample size condition and as the number of occasions increased, so it shouldn't be abandoned completely and should be evaluated further for LMI in longitudinal CFA.

Power was over 95% for all AFI's in both the large factor loading and large intercept LMI conditions. This is an indication that the values chosen may have been too large and any method a researcher chooses to use would easily detect a failure of MI this extreme.

Interestingly, the most commonly used methods for evaluating MI, besides the difference in chi-square, did not perform well. The decrease of greater than .01 in the CFI between models and the RMSEA C.I. criteria severely lacked power to detect LMI in the small loading LMI conditions. Specifically, power was 13% for the CFI at two time points and decreased to 0% by the sixth when $N = 400$. Power was lower for the larger N condition ($N = 800$). The RMSEA also fared poorly when $N = 400$ and although larger sample size increased power, it still fell well below the 80% recommendation and again decreased as measurement occasions increased. Here we have evidence that power does not always increase as repeated measurements increase.

For the small intercept conditions power was sufficient for two occasions of measurement using the CFI of .01, however, power decreased below 80% at the fourth measurement occasion and beyond. Once the number of indicators in the model were more than 28 and 42 for the small and large sample size conditions, respectively, there was not enough power to detect LMI. The more conservative change in CFI of .002 retained sufficient power across conditions. The RMSEA CI also had sufficient power until the sixth occasion at the smaller sample size.

Appropriate Null Model. The use of Widaman and Thompsons appropriate null model did not influence the evaluation of MI models using either of the CFI change criteria (.01 or .002). It does, however, influence the calculation of the CFI for a given model and may influence whether or not an initial model that is tested before evaluating MI is retained or rejected.

Number of Occasions. Type I error for the change in Gamma hat decreased as number of timepoints increased and this pattern was apparent in both sample size conditions (see Table 2). For the two most common AFIs cited in MI literature (CFI change and RMSEA C.I. comparisons) Table 2 shows the power to detect the small level of variant factor loadings decrease as occasion conditions increased for the CFI change of .01 and the RMSEA C.I. criteria. However, both of these AFIs has very low power across all conditions.

Power to detect the small level of variant intercepts dropped below 80% after the fourth occasion (N= 400) for the RMSEA C.I. criterion and after the second occasion for the change in CFI of .01 criterion. In the larger sample conditions (N = 800) the power of the CFI dropped at the sixth occasion condition. Number of time points did not seem to influence the power of the chi-square difference, change in gamma hat, nor the change in CFI of .002 criteria for intercepts or loading conditions.

Fan and Sivo (2007) found that the RMSEA and to a lesser extent the CFI both are influenced by different types of models. Specifically, the number of indicators in a model positively influences these indices. They also determined that Gamma hat was not influenced by the number of indicators since the formula accounts for number of items. The number of indicators increases exponentially in longitudinal CFA which may account for the drop in power

to detect LMI in this study after the fourth occasion. Although the CFI of .002 criterion was sufficiently powered.

Limitations

Perhaps the biggest limitation to the present study was the values used for large LMI conditions in both the loadings and the intercepts. Although a previous study was used as guidance, a range of values could provide more enlightenment on sizes of LMI. Additionally, significant LMI and practical LMI must be evaluated and may depend based on the measurement tool used and should also be considered.

An additional limitation is the relatively small number of conditions evaluated and the small number of items used to measure the latent construct. If a researcher is evaluating MI on a construct with over 20 or more items, power may not be sufficient for even two or three occasions using the CFI or RMSEA criteria since there is evidence that the number of indicators influence these AFIs. Essentially, once there are a certain number of indicators in a model, these AFIs do not move much and show undeserving good model fit.

As with any study, these findings only generalize to the limited conditions and population parameters laid out. Generalization or extrapolation beyond these is not recommended and further research and replication should be evaluated. However, what has been provided is evidence that researchers should not utilize rigid cutoffs for all types of CFA measurement models while evaluating MI.

Future Directions

The number of items for each latent construct (7 items) and values for small and large non-invariance were taken from Kim and Wilson (2014) since the first goal is to mimic those conditions. Future simulation studies can implement a range of values and additional conditions. Additionally, only up to eight repeated measurement occasions were examined in this study. Current technology allows researchers to collect many more occasions with a simple smart phone application leading to more occasions than CFA or SEM may be able to manage. Evaluation of MI in these scenarios will need to be investigated.

Parsimony error is common in CFA and SEM since the goal of these techniques is to find the most parsimonious (simplest) model to answer the research question of interest. The current study did not include any model misspecifications for parsimony error, which is why the chi-square difference tests performed so well. Future research should include parsimony error along with other potential model misspecifications in order to ascertain the affects on the AFIs ability to still detect LMI.

Asparouhov and Muthén (2014) developed the alignment method for simplifying MI testing in multiple-group CFA. However, it is only available in Mplus and cannot currently be specified for longitudinal CFA. There are other potential methods for testing MI such as determining appropriate effect sizes based on areas of study, or other approaches to hypothesis testing such as Bayesian or Neyman-Pearson. To my knowledge, none of these methods have been applied to longitudinal MI evaluation.

Conclusion

If researchers are interested in studying people overtime and conducting longitudinal studies, then this work is imperative. It is essential that researchers evaluate latent longitudinal

models for MI. Further research on how to evaluate longitudinal MI is needed. This study has shown the dangers of utilizing criteria based on cross-sectional to evaluate longitudinal MI by demonstrating that the most commonly used AFIs may not have adequate power to detect variant items. Therefore, even when researchers test for MI they may conclude that the measure is invariant over time. Thus, they will find differences over time and attribute them to true differences, rather than measurement invariance. These misinterpretations could lead researchers to conclude differences over time exist when actually the psychometric properties of the measurement tool differ but was not detected.

References

- Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in Quality of Life of People with Stroke over Time: True Change or Response Shift? *Quality of Life Research, 14*(3), 611–627.
- Askildsen, J. E., Jirjahn, U., & Smith, S. C. (2006). Works councils and environmental investment: Theory and evidence from German panel data. *Journal of Economic Behavior & Organization, 60*(3), 346-372.
- Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21*(4), 495-508.
- Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–108). New York: Guilford Press.
- Barclay-Goddard, R., Lix, L. M., Tate, R., Weinberg, L., & Mayo, N. E. (2009). Response shift was identified over multiple occasions with a structural equation modeling framework. *Journal of Clinical Epidemiology, 62*(11), 1181-1188.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.
- Bollen, K. A. (1989). *Structural equations with latent variables*. Hoboken, NJ: Wiley & Sons.
- Bollen, K. A., & Hoyle, R. H. (2012). Latent variables in structural equation modeling. In J.R. Harring & G.R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp.56-67). Thousand Oaks, CA: Sage.
- Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201-236).

- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. *Handbook of structural equation modeling*, 361-379.
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24), 4279-4292.
- Byrne, B. M. (1988). Measuring adolescent self-concept: Factorial validity and equivalency of the SDQ III across gender. *Multivariate Behavioral Research*, 23(3), 361-375.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.
- Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.
- Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, 1, 421-483.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.
- Chou, C. P., & Bentler, P. M. (2002). Model modification in structural equation modeling by imposing constraints. *Computational statistics & data analysis*, 41(2), 271-287.

- Chou, C. P., & Huh, J. (2012). Model modification in structural equation modeling. In J.R. Harring & G.R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp.232-246). Thousand Oaks, CA: Sage.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *Journal of abnormal psychology*, *112*(4), 558.
- DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137-149.
- Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149.
- Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532.
- Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531. <https://doi.org/10.1037/a0031669>
- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378-402.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144.

- Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review, 11*, 354-373
- Hoyle, R. H. (2012). Introduction and overview. *Handbook of structural equation modeling*, (pp. 3–16). New York: Guilford Press.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486-507.
- Kenny, D. A., & Milan, S. (2012). Identification: A non-technical discussion of a technical issue. *Handbook of structural equation modeling*, 145-163.
- Kim, E. S., & Willson, V. L. (2014). Testing Measurement Invariance Across Groups in Longitudinal Data: Multigroup Second-Order Latent Growth Model. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 566–576.
<https://doi.org/10.1080/10705511.2014.919821>
- King, K. M., Patock-Peckham, J. A., Dager, A. D., Thimm, K., & Gates, J. R. (2014). On the mismeasurement of impulsivity: Trait, behavioral, and neural models in alcohol research among adolescents and young adults. *Current Addiction Reports, 1*(1), 19-32.
- King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA Advances in Statistical Analysis, 94*(2), 139-156.
- King-Kallimanis, B. L., Oort, F. J., Nolte, S., Schwartz, C. E., & Sprangers, M. A. (2011). Using structural equation modeling to detect response shift in performance and health-related

- quality of life scores of multiple sclerosis patients. *Quality of Life Research*, 20(10), 1527-1540.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. (4th ed.). New York, NY: Guilford Press.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.
- Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. *Modeling contextual effects in longitudinal studies*, 121.
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, 13(1), 59–72.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9(2), 147.
- McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 37(2), 234-251.
- Mitchener, W. & Nowak, M. (2004). Chaos and language. *Proceedings: Biological Sciences*, 271(1540), 701-704.
- McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological methods*, 14(2), 126.
- McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods*, 7(1), 64.

- 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.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Millsap, R. E. & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York: Guilford Press.
- Mundform, D. J., Schaffer, J., Kim, M. J., Shaw, D., Thongteeraparp, A., & Supawan, P. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods, 10*(1), 4.
- O'Donnell, M. (2016). Is willingness to pay for non-consumptive wildlife watching falling? Evidence from three rounds of the national survey of fishing, hunting, and wildlife-associated recreation. *Human dimensions of wildlife, 21*(6), 475-490.
- Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*(3), 599-609.
- Pornprasertmanit, S., Miller, P., & Schoemann, A. (2016). *simsem* (Version 0.5-13).
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90.

- R Core Team. (2016). R: A language and environment for statistical computing (version 3.3.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from the comprehensive R archive network (CRAN): <https://www.R-project.org/>
- Raykov, T., & Marcoulides, G. (2011). *Introduction to Psychometric Theory*. New York: Taylor & Francis.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. <https://doi.org/10.1037/a0029315>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retrieved from <http://www.jstatsoft.org/v48/i02/>
- Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222.
- Schwartz, C. E., & Sprangers, M. A. . (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine, 48*(11), 1531–1548. [https://doi.org/10.1016/S0277-9536\(99\)00047-7](https://doi.org/10.1016/S0277-9536(99)00047-7)
- Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology & Health, 19*(1), 51–69. <https://doi.org/10.1080/0887044031000118456>
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.
- Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research, 25*(1), 78-90.

- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, 25(2), 173-180.
- Thompson, M. P., Sims, L., Kingree, J. B., & Windle, M. (2008). Longitudinal associations between problem alcohol use and violent victimization in a national sample of adolescents. *Journal of Adolescent Health*, 42(1), 21-27.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1), 4-70.
- Villa, K. M. (2017). Multidimensional human capital formation in a developing country: Health, cognition and locus of control in the Philippines. *Economics & Human Biology*, 27, 184-197.
- Wall, M. M., Guo, J., & Amemiya, Y. (2012). Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivariate Behavioral Research*, 47(2), 276–313.
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York: Guilford Press.
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Development Perspectives*, 4(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention:*

Methodological advances from alcohol and substance abuse research, 281-324.

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