PERFORMANCE OF SECOND-ORDER LATENT GROWTH CURVE MODELS WITH SHIFTING INDICATORS: A MONTE CARLO STUDY

by

KATHERINE ANNE RACZYNSKI

(Under the Direction of Deborah Bandalos)

ABSTRACT

Second-order latent growth models with shifting indicators (“the shifting indicators model”) allow longitudinal researchers to add or drop items to develop models that closely represent prevailing developmental theory. To date, however, published research evaluating the performance of the shifting indicators model has been minimal. Simulation methods were used to generate data where all indicators were present at all time points. Data for selected indicators were then deleted to create models with shifting indicators. The performance of shifting indicators models was compared to the original model with all indicators present. The number of shifting indicators per factor, the number of measurement occasions with shifting indicators, the magnitude of the factor loadings of the shifting indicators, and sample size was manipulated. Samples were drawn from multivariate normal populations, and for each cell 1000 replications were obtained. The results of the study indicated that the performance of the shifting models was quite similar to the performance of the models with all items included at each time point. Nonconvergence and inadmissible solutions were rare. Mean values of relative bias in the growth parameter estimates and their standard errors did not exceed .05 and .1, respectively, for all cells with sample size exceeding 250. The shifting indicators models were slightly less efficient than the models with all indicators present, but the difference was small. The investigation into model fit presented one potential caution. Having fewer items per factor was associated with better measures of fit, especially when the sample size was 250. This result indicates that model fit, as measured by chi-square and related fit indices, may be improved simply by dropping items from the model. The overall findings were promising and support the continued study of the shifting indicators model. A demonstration of the shifting indicators model with real data reinforced these findings.

INDEX WORDS: Second-order latent growth curve models, Shifting indicators, Monte Carlo, Longitudinal methods

PERFORMANCE OF SECOND-ORDER LATENT GROWTH CURVE MODELS WITH SHIFTING INDICATORS: A MONTE CARLO STUDY

by

KATHERINE ANNE RACZYNSKI

B.S. Ed., The University of Georgia, 2002

M.A., The University of Georgia, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

PERFORMANCE OF SECOND-ORDER LATENT GROWTH CURVE MODELS WITH SHIFTING INDICATORS: A MONTE CARLO STUDY

by

KATHERINE ANNE RACZYNSKI

Major Professor:	Deborah Bandalos
Committee:	Seock-Ho Kim Karen Samuelsen Arthur Horne

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

DEDICATION

This work is dedicated to my amazing family, who has supported me through everything.

ACKNOWLEDGEMENTS

I would first like to acknowledge and thank my wonderful advisor, Dr. Deborah Bandalos, who has been so instrumental and supportive throughout my entire graduate school career, and especially through this process of envisioning, conducting, and writing a dissertation. Thank you for many hours of guidance, instruction, and encouragement. You are an impressive scholar and an admirable role model. I also thank Drs. Seock-Ho Kim and Karen Samuelsen for serving on my committee and for many years of outstanding instruction. I have relied on your expertise throughout this process, and you have always been eminently generous on this count. My work is strengthened by having the opportunity to watch and learn from you.

There are a large number of people who have helped me develop professionally and personally. I would like to single out two mentors that I am especially grateful to: Drs. Pamela Orpinas and Andy Horne. Pamela, throughout our many years of working together, I have looked up to you for your command of the field, your admirable organizational and managerial skills running a large and complicated grant, and your unwavering dedication to producing excellent work. I continue to learn from you and am honored to call you a friend. Andy, you have always believed in me and pushed me take on greater and greater challenges, while providing me with the learning opportunities to help me get there. I look up to your example—as both an outstanding scholar and valued friend—and look forward to many more years of collaboration.

Finally, I would like to acknowledge the contributions of my wonderful family. Kevin, your love and support have been unflagging over the many sacrifices of graduate school. Mom

and Dad, thank you for providing me with a rock-solid foundation and encouraging me in a lifetime love of learning. To Leslie, J, Caroline, Anna, and the rest of my family and friends, thank you for your patience and encouragement as I have finished this degree. I could not have done this without you.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Investigating change is a fundamental component of empirical research. Rather than merely collecting a snapshot of the targeted phenomenon, psychological and behavioral researchers are often interested in drawing inferences about how and why outcomes change over time. Fortunately, many options exist for researchers who want to move beyond cross-sectional studies. In recent years, longitudinal research methods have become more widespread, more sophisticated, and more accessible (Singer & Willett, 2003, p. 3).

Two common overarching goals of longitudinal data analysis are investigating changes in means over time and changes in individual differences over time (Marsh & Grayson, 1994). Take, for example, a team of researchers who are interested in measuring bullying behaviors during childhood. The team may want to understand the mean amount of bullying that takes place over time. On average, do rates of bullying increase, decrease, or stay the same as students get older? Does growth or decline take a linear form, or does it follow some other pattern? The team likely also wants to investigate differences in individuals. Do individuals start with about the same amount of aggressive behavior? Do they follow the same developmental trajectories over time? Are differences in starting points related to differences in trajectories?

There are several ways to investigate these types of questions. In this study, the primary focus is on latent growth curve models (LGMs), a structural equation modeling (SEM) based approach to longitudinal data analysis. LGMs are a popular approach to investigating longitudinal change due to their flexibility and advantages over methods of longitudinal data analysis such as MANOVA and lagged regression. Lance, Vandenberg, and Self (2000) summarize these advantages, which include the ability to (a) model individual and group-level change, (b) describe individual differences in slope and intercept, (c) investigate change at the latent-construct level (i.e., accounting for measurement error), (d) investigate different types of growth trajectory (e.g., linear, quadratic), (e) model growth in multiple constructs simultaneously (i.e., multivariate growth modeling), (f) include predictors and outcomes related to intercept and slope, and (g) test hypotheses related to mediators of longitudinal change. Other longitudinal analysis methods may include some but not all of these advantages (Lance, Vandenberg, & Self, 2000).

Despite these advantages, the most popular type of LGMs—first-order LGMs—have several drawbacks. First-order LGMs include a single indicator (i.e., manifest variable) at each time point. This indicator is typically a sum or average of several items (Ferrer, Balleurka, & Widaman, 2008). One problem is that by utilizing a single composite variable, restrictive assumptions are made regarding the relationships of the items to each other and to the latent construct. Specifically, items are assumed to relate to the latent construct in the same way (i.e., equal factor loadings) as each other and over time. Another drawback is that the measurement errors of individual items are incorporated into the composite (Hancock, Kuo, & Lawrence, 2001). Further, first-order LGMs lack a mechanism for evaluating longitudinal measurement invariance.

Leite (2007) demonstrated that LGMs that use composite scores yielded adequate results only under restrictive conditions. To meet these conditions, the items of which the scale is composed needed to have equal loadings with each other and demonstrate strict longitudinal invariance (i.e., equal factor loadings, item intercepts, and unique variances over time). These conditions are not likely to be encountered often in applied settings. Complicating matters, longitudinal measurement invariance cannot be explicitly tested in first-order LGMs, although it is a prerequisite for investigating change. Therefore, first-order LGMs utilizing composite scores may not be the most appropriate method for examining growth.

Second-order LGMs provide a more comprehensive way to understand growth that leverages several advantages over first-order LGMs. Second-order LGMs incorporate a measurement model into the examination of growth; that is, the latent construct of interest is measured by multiple indicators at each time point (Hancock, Kuo, & Lawrence, 2001). Each item has an associated error term, and the latent construct at each time point has a disturbance term which captures occasion-specific (i.e., transient) error. The result is that the focus of the investigation—the structural model concerned with the growth of latent constructs over time—is rendered theoretically error-free (Hancock, Kuo, & Lawrence, 2001; Chan, 1998).

Second-order LGMs can be used to explicitly test measurement invariance assumptions that underlie investigations of longitudinal change, including measurement invariance at the configural (i.e., factor structure), metric (i.e., factor loading), and scalar (i.e., item intercept) levels. This investigation is achieved via a series of nested models; chi-square difference tests are employed to determine whether adding equality constraints on item loadings and intercepts across time results in significantly poorer model fit (Ferrer, Balleurka, & Widaman, 2008).

In standard second-order LGMs, identical indicators are present at each time point. Depending on the nature of the study and the construct measured, identical indicators may not always be available or appropriate. Hancock and Beuhl (2008) describe a variation of a second-order LGM where shifting indicators of the latent construct across time are utilized. In longitudinal studies that follow participants through developmental changes, appropriate indicators of the same latent construct may change throughout the study. For example, a study of the development of problem solving skills may include different measures for toddlers versus school-aged participants. The goal is to measure the same underlying construct while presenting the most relevant set of indicators at each time point. Presenting different but appropriate indicators across time is a particularly salient issue for research on developmental changes over the lifespan (McArdle & Grimm, 2011).

Measures may also undergo revision over time. Items that have not demonstrated adequate psychometric properties over time may be dropped and replaced with more appropriate ones. The wording of items may be updated to reflect changes in popular culture or nomenclature. For example, a set of items measuring cyberbullying may need to be revised relatively frequently to reflect rapidly-changing technologies.

Practical considerations may also limit the availability of the same indicators at each data collection wave. For example, funding limitations may impact the ability of researchers to collect full information at each assessment wave. Studies may also lack complete data when errors are made with data collection or if data become corrupted. In short, there are a variety of theoretical and practical reasons why a model that allows for shifting indicators may be beneficial to longitudinal researchers.

Despite these advantages, second-order LGMs—both with and without shifting indicators—have received relatively little attention in the literature, both in terms of applications of the model and methodological investigations of its performance. To my knowledge there is only one applied (Pettit, Keiley, Laird, Bates, & Dodge, 2007) and one methodological (Hancock & Buehl, 2008) examination of the performance of second-order LGMs with shifting indicators. In the applied study, parental monitoring was examined as students progressed from grade 5 to grade 11. As the children got older, the questions changed to account for the types of activities older children may engage in. A second-order LGM incorporating these shifting indicators was employed to investigate changes in parental monitoring over time.

In the methodological examination of second-order LGMs with shifting indicators, Hancock and Buehl (2008) demonstrated several desirable characteristics of the model. Using simulated data, the identical and shifting indicators models recovered the same solution, and when real data were used, similar results were obtained. However, the scope of this investigation was relatively limited. Only one real data set was used, and this data set demonstrated excellent fit and overall longitudinal invariance.

The purpose of this study is to further evaluate the performance of second-order LGMs with shifting indicators. Given the general lack of published research utilizing second-order LGMs with shifting indicators, longitudinal researchers in the social sciences may be unaware of these models or uncertain about when they may be appropriate. This study aims to provide a more fine-grained understanding of the performance of these models with special attention paid to sample size, the number of measurement occasions with shifting indicators, the proportion of shifting to non-shifting indicators, and the magnitude of the factor loadings of the shifting indicators. The next chapter provides a more comprehensive overview of LGMs and longitudinal measurement invariance and connects this literature to the research questions and rationale for the current study.

CHAPTER 2

REVIEW OF LITERATURE

Well-designed longitudinal studies aim to understand changes in the construct(s) of interest over time while minimizing the impact of irrelevant influences. Latent growth curve models (LGMs) are a technique for longitudinal analysis that have gained popularity for their flexibility, ease of use, and desirable measurement characteristics (Lance, Vandenberg, & Self, 2000). Second-order LGMs with shifting indicators are a potentially useful extension to traditional LGMs, but the performance of this type of model has not been extensively evaluated. The purpose of this study is to evaluate the performance of second-order LGMs with shifting indicators using Monte Carlo simulation.

In this chapter, I provide an overview and comparison of first-order LGMs, second-order LGMs, and second-order LGMs with shifting indicators. I describe how the use of shifting indicators of a construct over time may enhance the ability of researchers to investigate change given a variety of theoretical and practical considerations, and I present the research questions and rationale for the current study. I argue that second-order LGMs with shifting indicators represent a promising approach that may allow a larger number of researchers to employ a more sophisticated framework for assessing change, but because these models are relatively untested, additional research is warranted.

First-order LGMs

LGMs are a popular and flexible approach to investigating longitudinal change. LGMs combine aspects of variable-centered and person-centered analyses. Researchers may model linear or non-linear trajectories, and measurement occasions may be equally or unequally spaced. LGMs can include predictors and/or outcomes related to intercept and slope, among many other extensions to the model (Hancock & Lawrence, 2006).

In a first-order LGM, a single indicator variable is included in the model at each time point, typically a composite score such as a sum or average of multiple items (Ferrer, Balleurka, & Widaman, 2008; Leite, 2007). This type of LGM can be represented by the covariance and mean structures of the composites. The covariance structure is expressed as

$$\mathbf{\Sigma}_{cc} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}_{\varepsilon}, \tag{1}$$

where $\mathbf{\Sigma}_{cc}$ is the covariance matrix of the composites, $\mathbf{\Lambda}$ is the matrix of loadings relating the composites to the latent intercept and slope factors, $\mathbf{\Phi}$ is the covariance matrix of the latent intercept and slope factors, and $\mathbf{\Theta}_{\varepsilon}$ is the covariance matrix of measurement errors associated with the composites.

The mean structure of the composites is expressed as

$$\boldsymbol{\mu}_{cc} = \mathbf{\Lambda}\boldsymbol{\mu}_{\alpha\beta}, \tag{2}$$

where $\boldsymbol{\mu}_{cc}$ is the vector of the means of the composites, $\mathbf{\Lambda}$ is the matrix of factor loadings relating the composites to the latent intercept and slope factors, and $\boldsymbol{\mu}_{\alpha\beta}$ is the vector of the means of the latent intercept and slope.

In many configurations of LGMs, five key parameters are estimated to describe the data: the intercept mean (μ_{α}) and variance (ψ_{α}), the slope mean (μ_{β}) and variance (ψ_{β}), and the slope/intercept covariance ($\psi_{\alpha\beta}$). The means of the intercept and slope represent the average initial level and rate of change, respectively, across individuals. The variances of the intercept and slope describe how individuals vary with respect to starting point and rate of change.

Variance values that are significantly different from zero indicate that significant variation in individuals' intercepts and slopes exists in the data. The slope/intercept covariance provides an indication of the relationship between the intercept and the slope.

First-order LGMs make up the majority of LGMs employed in research. A first-order LGM with four waves of data is presented in Figure 2.1. Manifest variables (i.e., Y1-Y4) are represented by squares, and latent variables (i.e., intercept and slope) are represented by circles. All paths connecting the intercept to manifest variables are set to one, and the paths connecting the slope to each manifest variable are set to [0, 1, 2, 3] to invoke a linear growth pattern at four equally-spaced time points. At each time point, an error term associated with each manifest variable represents variance in the indicator that is not explained by the latent variables. The constant 1, represented by a triangle, is included to obtain the means of the intercept (μ_α) and slope (μ_β) factors. The other key parameters, namely the variances of the intercept (ψ_α) and slope (ψ_β) and the intercept/slope covariance ($\psi_{\alpha\beta}$), are also included in the model. A wide range of variation on this model is possible. For instance, LGMs may include non-linear growth trajectories, data collected at unequally spaced measurement occasions, and predictors of the slope and intercept, among many other options (e.g., Preacher, Wichman, MacCallum, & Briggs, 2008).

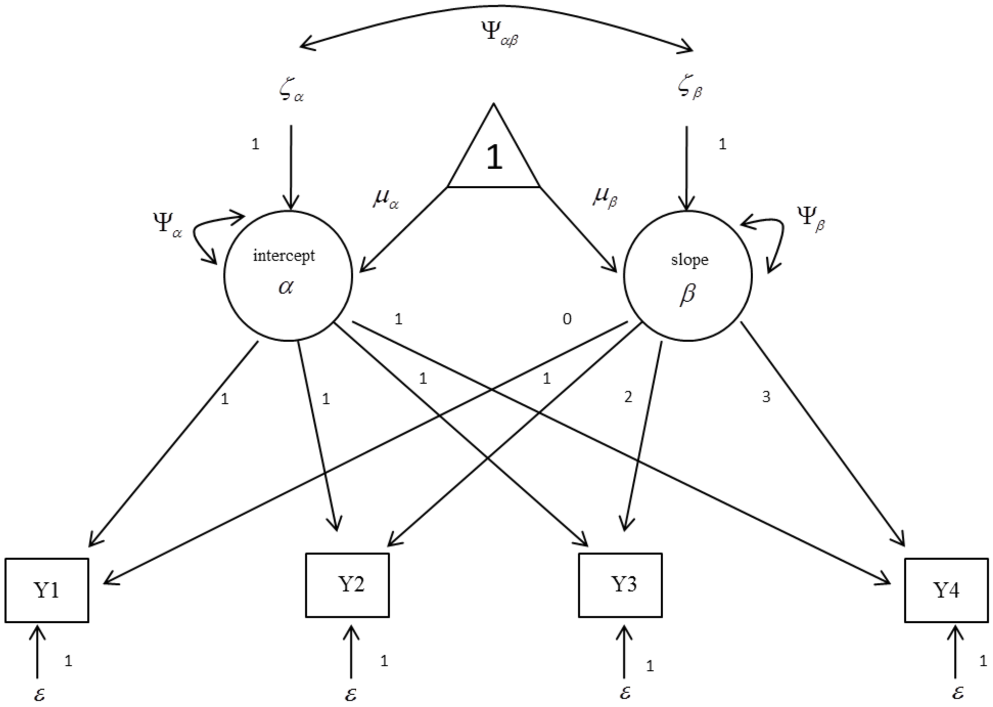


Figure 2.1. First-order LGM.

Second-order LGMs

A fairly recent extension of latent growth curve modeling involves incorporating a measurement model into the examination of growth. A second-order LGM investigates growth in a latent construct of interest measured by multiple indicators at each time point (Hancock, Kuo, & Lawrence, 2001). This type of model has also been called a curve of factors model (McArdle, 1988), a latent variable longitudinal curve model (Tisak & Meredith, 1990), and a multiple indicator growth model (Muthén & Muthén, 2010).

The second-order LGM incorporates a latent growth model, a measurement (i.e., factor analytic) model, and the mean structure of the items. The measurement model can be expressed as

$$\mathbf{\Sigma}_{yy} = \mathbf{\Lambda}_y \mathbf{\Omega} \mathbf{\Lambda}'_y + \mathbf{\Theta}_\varepsilon \quad , \tag{3}$$

where $\mathbf{\Sigma}_{yy}$ is the covariance matrix of the items, $\mathbf{\Lambda}_y$ is the matrix of loadings relating the items to the latent constructs, $\mathbf{\Omega}$ is the covariance matrix of the latent factors, and $\mathbf{\Theta}_\varepsilon$ is the covariance matrix of the item uniquenesses. The covariance matrix of the latent factors is expressed as

$$\mathbf{\Omega} = \mathbf{\Lambda}_\eta \mathbf{\Phi} \mathbf{\Lambda}'_\eta + \mathbf{\Psi} \, , \tag{4}$$

where $\mathbf{\Lambda}_\eta$ is the covariance matrix of the loadings relating the latent constructs to the latent intercept and slope, $\mathbf{\Phi}$ is the covariance matrix of the intercept and slope, and $\mathbf{\Psi}$ is the covariance matrix of the disturbances (ζ).

The mean structure of the items includes a first-order component relating the item means to the latent construct means, and a second-order component relating the latent construct means to the means of the latent intercept and slope. The mean structure relating the item means to the latent construct means is

$$\boldsymbol{\mu}_y = \mathbf{\Lambda}_y \boldsymbol{\mu}_\eta + \boldsymbol{\tau}_y \, , \tag{5}$$

where $\boldsymbol{\mu}_y$ is the vector of item means, $\mathbf{\Lambda}_y$ is the covariance matrix of the loadings relating the latent constructs to the latent intercept and slope, $\boldsymbol{\mu}_\eta$ is the vector of means of the latent constructs, and $\boldsymbol{\tau}_y$ is the vector of item intercepts.

The mean structure relating the latent construct means to the latent intercept and slope is

$$\boldsymbol{\mu}_{\eta} = \boldsymbol{\Lambda}_{\eta}\boldsymbol{\mu}_{\alpha\beta} , \tag{6}$$

where $\boldsymbol{\Lambda}_{\eta}$ is the covariance matrix of the loadings relating the latent constructs to the latent intercept and slope and $\boldsymbol{\mu}_{\alpha\beta}$ is the vector of the means of the latent intercept and slope.

Figure 2.2 presents a second-order LGM with four waves of data measured at equally spaced time points. Compared with a first-order LGM, the model presented in Figure 2.1 adds a second level of latent variables (η) representing a latent construct measured by multiple items at each time point (in this case, grades 6 – 9). The factor loading (λ) for each item is interpreted as a regression slope relating the observed score to the latent construct. That is, loadings represent the amount of change in the observed score given a one unit change in the amount of the latent construct. Item intercepts (τ) represent the value of the observed score on an item when the value of the latent construct is zero.

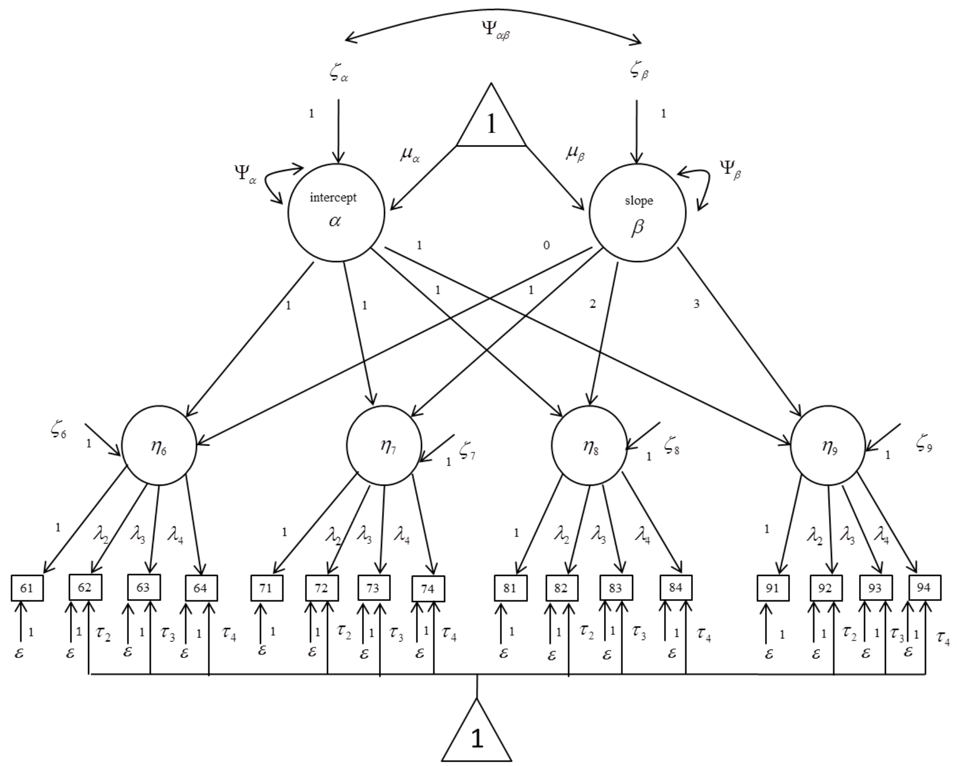


Figure 2.2. Second-order LGM with identical indicators.

In order to identify the measurement part of the model, one indicator at each time point is designated the scale indicator. The loadings of the scale indicators are set to one and the intercepts are set to zero. This sets the metric of the latent variable equal to the metric of the scale indicator. Intercepts for the remaining indicators are obtained by regressing the variables on the constant 1 (represented by a triangle). In the example presented in Figure 2.2, intercepts and loadings for corresponding items (e.g., item 2 measured at each time point) are constrained to be equal.

Advantages of Second-order LGMs

Second-order latent growth curve modeling provides several advantages over first-order latent growth curve modeling. These advantages relate to the use of multiple indicators at each time point rather than a composite (e.g., mean) score. As stated earlier, first-order LGMs include only one indicator, often the mean of several items, per measurement occasion. However, the use of a composite score introduces several potential sources of bias. First, a mean score does not differentially weight the items composing the scale (Hancock, Kuo, & Lawrence, 2001). Each item contributes equally to the mean, which implies that each item is equally related to the construct. Second-order LGMs, on the other hand, are used to directly model the growth of the latent constructs rather than the observed indicators. By employing a latent variable measurement model, items are weighted according to their relationship to the construct; items that are more related to the latent construct (i.e., have higher factor loadings) are weighted more heavily (Sayer & Cumsille, 2001). Further, the unique variance for each indicator (ε) is explicitly modeled instead of being incorporated into a composite (Sayer & Cumsille, 2001). This error term represents a combination of random response error and item-specific errors of measurement. Under first-order LGMs these errors are confounded with transient (i.e., occasion-specific) error.

A second problem with the use of composite scores is the inability to explicitly test for measurement invariance. Measurement invariance refers to a test's ability to measure the same latent variable under different conditions, such as at different measurement occasions (Horn & McArdle, 1992). Before interpreting results obtained from a common measure over time, researchers should evaluate whether respondents respond to and interpret the measure in a similar way across measurement occasions. Without evidence of longitudinal measurement invariance, differences in item response over time may be due to actual growth or decline in the construct of interest or to differences in how respondents interact with the measure.

One process of testing for measurement invariance involves comparing the fit of nested models as equality constraints are placed on item parameters in a stepwise fashion. The most common levels of measurement invariance investigated are configural (i.e., pattern of fixed and free factor loadings), metric (i.e., equality of factor loadings), and scalar (i.e., equality of item intercepts). The equality of the unique variance of corresponding items across time may also be investigated. Measurement invariance is supported when adding equality constraints does not result in significantly worse model fit as measured by the chi-square difference test (Vandenberg & Lance, 2000).

When individual items are collapsed into a composite score, the relationship of each item to the latent construct is lost, and it is not possible to conduct this series of tests. Therefore, in most cases, longitudinal invariance is assumed, but not investigated (Lance, Vandenberg, & Self, 2000; Ferrer, Balleurka, & Widaman, 2008). However, when measurement invariance does not hold or is left untested, real differences in the construct over time may be confounded with measurement differences, and score interpretations, as well as interpretations of the growth parameters, may be flawed.

The problems with composite scores described above have important implications for the use of first-order LGMs. Specifically, first-order LGMs may perform well only under restrictive conditions. In a simulation study conducted by Leite (2007), first-order LGMs performed adequately only when (a), all items had equal loadings in relation to each other, (b), all items had equal loadings over time (i.e., demonstrated metric invariance), (c), all items had equal intercepts over time (i.e., demonstrated scalar invariance), and (d), all items had equal unique variances over time. Finding measures that meet this strict set of restrictions may be difficult to achieve in practice (Byrne & Stewart, 2006). Given second-order LGMs, some of these restrictions may be relaxed. First, the items that make up the measure do not necessarily need to have factor loadings that are equal to each other. In second-order LGMs, an individual loading is estimated for each item, and the items are not collapsed into a composite, thus eliminating the need for items with equal loadings.

Second, while the measure should exhibit an adequate degree of longitudinal measurement invariance, the minimum level of invariance to be met may be less stringent when second-order LGMs are employed. If the items that compose a measure are intended to be collapsed into a composite score, the measure should exhibit adequate measurement invariance at the configural, metric, scalar, and unique variance levels (Bontempo & Hofer, 2007; Schmitt & Kuljanin, 2008). However, when items are not combined, an adequate degree of invariance at the configural, metric, and scalar levels has been deemed sufficient (Meredith, 1993; Byrne & Stewart, 2006; Bontempo & Hofer, 2007). By explicitly modeling the unique variances of each of the items rather than collapsing them into a composite, the researcher does not need to meet the restrictive condition of equal unique variances over time.

Further, using a second-order LGM allows the researcher to try to establish partial measurement invariance if full measurement invariance is not met. If metric, scalar, or item uniqueness invariance does not hold, the researcher may conduct additional analyses to attempt to establish partial measurement invariance. A measure exhibits partial measurement invariance when a subset of item parameters is equivalent over time (Byrne, Shavelson, & Muthén, 1989). The equality constraints on the non-invariant parameters are released, and these parameters are estimated freely to improve model fit. There is some debate in the literature about whether and how partial measurement invariance should be included in latent variable models (e.g., Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000); however, it is clear that exploring partial measurement invariance in the absence of full invariance is a popular option for many researchers. As evidence of this popularity, a review of empirical studies using the term “measurement invariance” published from 2000 to 2007 indicated that 50% of studies included some test of partial invariance (Schmitt & Kuljanin, 2008). In sum, second-order LGMs allow researchers to investigate longitudinal change at the latent construct level, thereby avoiding many of the limitations and problems implicit with using composite scores.

Second-order LGMs with Shifting Indicators

While traditional LGMs employ the same items at each time point, under some circumstances it may be desirable in longitudinal research to drop or add items at different time points. Hancock and Beuhl (2008) describe an extension of second-order LGMs to incorporate changing indicators of the latent construct over the course of the study. While the latent construct is theorized to remain conceptually the same, the items used as indicators of the construct are allowed to vary across time.

Figure 2.3 depicts a second-order LGM with shifting indicators that is a reduced form of the model presented in Figure 2.2. Data are still collected in four waves, but the second item has been dropped from the sixth-grade wave of data (i.e., item 62), and the fourth item has been dropped from the ninth-grade wave of data (i.e., item 94). This example is meant to reflect a situation where one item (i.e., item 62) is not developmentally appropriate at the first time point, and one item (i.e., item 94) is not developmentally appropriate at the last time point. The first item functions as the scale indicator with the loading set to one and the intercept set to zero at each time point. As with the full indicator model presented in Figure 2.2, other loadings and intercepts are set to be equal across corresponding items to reflect the case where longitudinal measurement invariance holds.

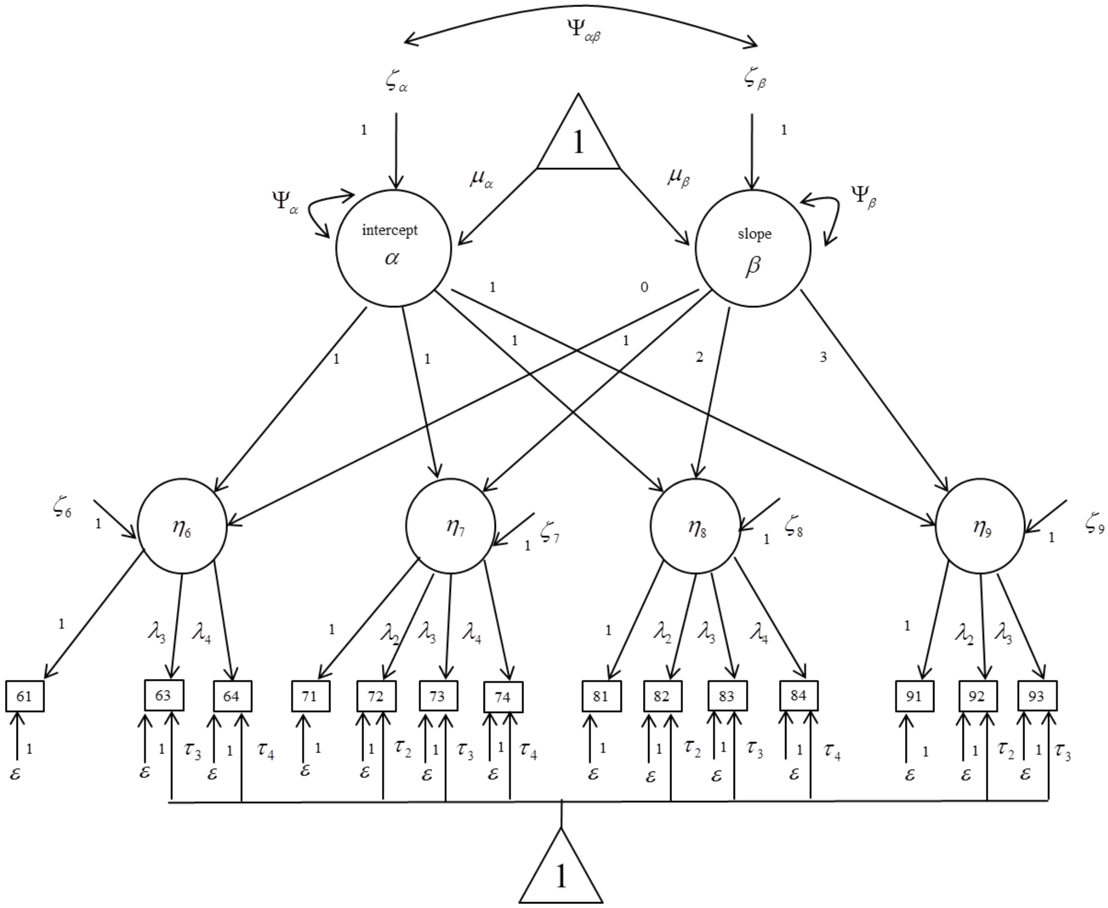


Figure 2.3. Second-order LGM with shifting indicators.

Scaling of Second-order LGMs with Shifting Indicators

As with the full indicator model, a scale indicator may be used in the shifting indicators model to set the metric for the measurement model. (Note that “the shifting indicators model” and “the shifting model” will be used as shortened terms for “second-order LGM with shifting indicators” in the remainder of this manuscript.) With the shifting indicator model, it is possible to set the scale of the measurement model even when the scale indicator is not present at each time point. Hancock and Buehl (2008) presented an example where a total of five items are measured over four time points, but there was little overlap among items. In this study, the five items were labeled with letters (i.e., A-E) instead of numbers. Time 1's indicators were A and B, time 2's indicators were B and C, time 3's indicators were C and D, and time 4's indicators were D and E.

The scale of the measurement model was set as follows. Item A was used as the scaling indicator, although it only was only measured at time 1. All of the corresponding loadings and intercepts for the other items were set to be equal over time. For example, the loadings and intercepts for item B were set to be equal at times 1 and 2, the loadings and intercepts for item C were set to be equal at times 2 and 3, and so forth. Because of the equality constraints on the loadings, the items were linked across time, and the scaling set by item 1 at the first time period was carried forward to the other time points. Specifically, at time 1, a one-unit change in the latent variable (i.e., η_1) resulted in a one-unit increase in the value of item A and λ_B-unit increase in the value of item B. At time 2, a one-unit change in η_2 also resulted in a λ_B-unit increase in the value of item B, as well as a λ_C-unit increase in the value of item C. At time 3, a one-unit change in η_3 resulted, again, in a λ_C-unit increase in the value of item C along with a λ_D-unit increase in the value of item D. At time 4, a one-unit change in η_4 resulted in a λ_D-unit increase in the value of item D and a λ_E-unit increase in the value of item E. The connection between each set of corresponding items allowed for the entire set of measures to be placed on the same scale.

Performance of Second-order LGMs with Shifting Indicators

Second-order LGMs with shifting indicators have performed well in evaluations of the performance of the model, although few such studies have taken place. Hancock and Beuhl (2008) used one simulated and one real dataset to demonstrate the flexibility of the shifting indicators model. As described above, they present models where the measured indicators at each time point are common to adjacent time points alone. Although these models had a minimal degree of item overlap, they reported that the performance of the models was comparable to corresponding full models with all items present at each time point (i.e., “the full model”). Under simulated data, the full model and the shifting indicators models recovered the same correct estimates of key parameters. In the set of analyses using real data, some variation occurred in parameter estimation, but the pattern of significant and non-significant values of key parameters was the same in the two models (Hancock & Beuhl, 2008). Specifically, under the full and the shifting indicators models, the intercept mean and variance were significant, all other key parameters were non-significant, and the comparative fit index (CFI) and root mean square error of approximation (RMSEA) values indicated adequate fit. These results are promising, as there may be theoretical or practical reasons for longitudinal researchers to drop or add indicators of a latent construct over time.

Advantages of Second-order LGMs with Shifting Indicators

Second order LGMs with shifting indicators represent an approach to data analysis that may provide solutions to some common problems in longitudinal data collection and analysis. A promising application of second-order LGMs with shifting indicators involves the use of the models to account for theoretically expected changes in the manifestation of a latent construct over time. Depending on the nature of the construct, appropriate indicators of a latent construct of interest may shift in predictable ways as participants progress through developmental stages. For instance, in a study of the development of aggression in children, an item that asks about stealing other children’s toys may no longer be relevant after a certain age. At the same time, other indicators of aggression may become more appropriate, even as the overall construct of aggression is conceptually unchanged. In this situation, theory dictates which indicators are appropriate at a given stage and why. Given a well-developed theory and high quality indicators to choose from, researchers have the ability to use second-order LGMs with shifting indicators to develop models that closely represent prevailing developmental theory.

When longitudinal research is conducted over a long period of time, the ability to add or drop indicators may become especially valuable. Consider the usefulness of an indicator such as alcohol use as a measure of deviancy from childhood through early adulthood. In an elementary age population, one would imagine alcohol use to be quite rare, even among students with high levels of deviancy. Deviancy at this age tends to be associated with, for example, fighting, stealing, and lying, rather than drug use. In middle and high school, alcohol use may be a more appropriate indicator of deviancy, yet in early adulthood (i.e., after age 21), moderate alcohol use ceases to be a viable indicator of deviancy, given that it becomes a legal activity. Assuming that theory supports a common definition of deviancy across this time span, a shifting indicators model can measure the trajectory of deviancy while accommodating these changes in the manifestations of the latent construct.

Above and beyond theoretical reasons, practical considerations may contribute to the need to use shifting indicators. Changes in cultural touchstones and popular nomenclature may predicate changes in the measurement of a construct over time. As norms change, existing items may no longer be appropriate. A measure of job satisfaction, the Job Description Index, was developed in 1969 and has been revised several times based on societal changes in the workplace and common phrasing regarding work (Lake, Gopalkrishnan, Sliter, & Withrow, 2010). Chan, Drasgow, and Sawan (1999) investigated the performance of items on the Armed Services Vocational Aptitude Battery (ASVAB) and found that the item parameters of 25 out of 200 items changed significantly over a period of 16 years. Items that were more semantically laden were more likely to exhibit measurement changes, indicating that shifts in the use of language over time may contribute to differences in item response. Some fields, such as those related to consumer technology or popular culture may require relatively frequent updates to ensure that items are relevant.

Other unexpected occurrences may render a longitudinal dataset incomplete. For instance, administrative or printing errors may result in items being inadvertently dropped from the dataset. A reduction in funding or access to participants may cause researchers to exclude some indicators from certain waves of data collection. Data may be incorrectly gathered in the field or corrupted before analyses take place. All of these problems represent less than ideal circumstances impacting the availability of indicators at each time point. However, these problems are not rare. Given sufficient overlap in indicators and evidence that the intact data are suited for longitudinal analysis (e.g., exhibit longitudinal invariance), second-order LGMs with shifting indicators may provide an opportunity to salvage a line of research in light of unexpected setbacks.

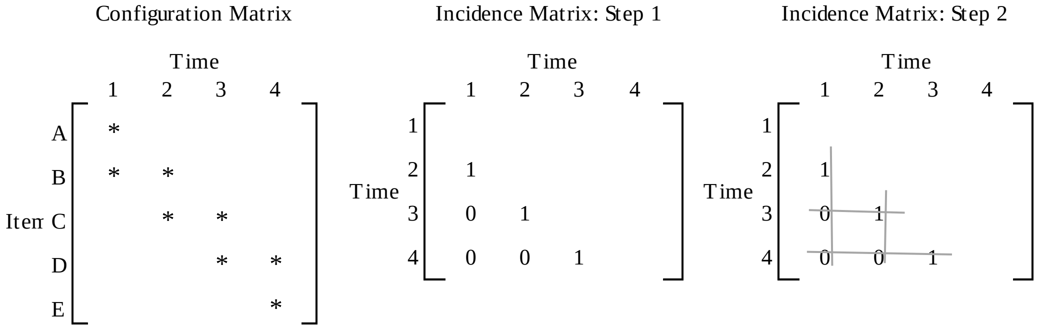
Theoretical and Practical Considerations Pertaining to the Use of Second-order LGMs with Shifting Indicators

While second-order LGMs with shifting indicators may provide solutions to some measurement problems faced by researchers, they are not appropriate in all situations and should not be employed in the absence of sound theory. As mentioned earlier, before engaging in any type of latent growth curve modeling, the latent construct of interest should be stable in meaning across time. Manifestations of the construct may change across time, but the essence of the construct should, as defined by theory, remain the same. For example, the indicators of parental monitoring may be different in middle school than in high school (where one would expect high school students to have more freedom), but the underlying construct likely remains the same. It is possible that theory dictates the essence of the latent construct *does* change. For instance, theory may suggest that parental monitoring in infancy (e.g., attending to and being receptive of the baby's needs) is qualitatively different than monitoring in adolescence (e.g., knowing the child's friends, following the child's progress in school). In this case, latent growth modeling of monitoring as a single construct across time would be inappropriate.

Given that there is a theoretical basis for the construct's stability in meaning across time, theory should also inform the inclusion or exclusion of indicators (given that items are to be added or dropped in a planned fashion). Theory must be detailed enough to guide the selection of appropriate indicators at each time point. Researchers should be able to explain the theoretical basis for the inclusion or exclusion of each item at each measurement occasion. This may prove to be a stumbling block for some researchers, as the theory may not be sufficiently well developed to provide direction at this level of detail.

In addition to theoretical preconditions, several practical considerations impact the appropriateness of second-order LGMs with shifting indicators. First, there must be sufficient overlap across indicators to identify the model. Hancock and Buehl (2008) provide guidance for determining whether sufficient linkages exist across time. The first step involves developing a configuration matrix. In this matrix, asterisks indicate at which time points each item is measured. From the configuration matrix, an incidence matrix is created. The incidence matrix indicates which time points have one or more constrained indicators in common. Time points with shared indicators are designated with a "1," and time points with no shared indicators are designated with a "0." In order to meet the minimum amount of overlap needed for model identification, the incidence matrix must have a minimum of *T*-1 non-zero elements (i.e., "1's") below the diagonal arranged in a particular configuration, where *T* is the number of measurement occasions. Vertical and horizontal lines are drawn through the row or column containing each non-zero element below the diagonal. These lines are extended to cross out the elements above, below, and to the sides of the non-zero elements. If all of the elements are crossed out after the lines are extended, the minimum condition for model identification has been met.

A model presented by Hancock and Buehl (2008) is used to illustrate this process. The relevant matrices are displayed in Figure 2.4. As described earlier, the model involves four time points and five items. From the configuration matrix, overlap among the items across time points can be easily identified. In this case, there is overlap among time 1 and time 2 (i.e., item B), time 2 and time 3 (i.e., item C), and time 3 and time 4 (i.e., item D). This information is used to develop the incidence matrix. In the second step involving the incidence matrix, vertical and horizontal lines are drawn through the non-zero elements of the incidence matrix. In this example, all of the zero elements are crossed out by the lines, indicating that the minimum amount of overlap for model identification is met.



Configuration Matrix

$$\begin{array}{c|cccc} & \multicolumn{4}{c}{\text{Time}} \\ & 1 & 2 & 3 & 4 \\ \hline \text{A} & * & & & \\ \text{B} & * & * & & \\ \text{Item C} & & * & * & \\ \text{D} & & & * & * \\ \text{E} & & & & * \end{array}$$

Incidence Matrix: Step 1

$$\text{Time}\ \begin{array}{c|cccc} & \multicolumn{4}{c}{\text{Time}} \\ & 1 & 2 & 3 & 4 \\ \hline 1 & & & & \\ 2 & 1 & & & \\ 3 & 0 & 1 & & \\ 4 & 0 & 0 & 1 & \end{array}$$

Incidence Matrix: Step 2

$$\text{Time}\ \begin{array}{c|cccc} & \multicolumn{4}{c}{\text{Time}} \\ & 1 & 2 & 3 & 4 \\ \hline 1 & & & & \\ 2 & 1 & & & \\ 3 & 0 & 1 & & \\ 4 & 0 & 0 & 1 & \end{array}$$

Figure 2.4. Matrices involved in determining sufficient overlap.

Some circumstances faced by researchers may not meet these model identification conditions. For example, if researchers want to use an entirely new set of indicators to measure the latent construct, there must be a wave of data collection where the new and at least some of the old measures are administered. Depending on the length of the measures and other factors, this may not be practicable. Further, in the case of unexpected loss or corruption of data, there may be insufficient intact indicators to identify the model.

Another practical consideration is that invariant indicators need to be available at each time point. If indicators that lack measurement invariance are selected, model fit will likely be unacceptable and interpretations of the outcomes will be flawed. Ferrer et al. (2008) demonstrated the consequences of using indicators that failed to meet invariance assumptions using second-order LGMs. Using a dataset that failed to exhibit longitudinal invariance at any level, the authors estimated a second-order LGM using each indicator, in turn, as the scale indicator. In this dataset, the choice of scale indicator greatly impacted the parameter estimates, resulting in different interpretations depending on the item designated as the scale indicator. As much as possible, researchers should select indicators that have already demonstrated favorable measurement characteristics, including exhibiting longitudinal invariance. However, this principle does not exclusively apply to the use of second-order LGMs with shifting indicators. Regardless of the type of analysis that is employed, the use of high-quality items is essential to drawing meaningful inferences.

Comparison of Second-order LGMs with Shifting Indicators to Item Response Theory Approaches

At a basic level, the goal of second-order LGMs with shifting indicators is to render comparable scores on measures that differ to some extent. Researchers operating under many of the theoretical traditions within the field of measurement (e.g., classical test theory, item response theory) have developed other techniques for approaching this kind of research goal. In this section, I describe several alternative approaches to obtaining comparable scores—both cross-sectionally and longitudinally—and compare them to second-order LGMs with shifting indicators.

By way of background, a brief introduction to item response theory (IRT) is first presented. IRT models relate item response to the level of a latent trait, θ, possessed by each respondent, and characteristics of the item. In a dichotomously scored item, the probability of endorsing the item (e.g., responding "yes" in a symptom checklist) increases as the level of the underlying latent trait increases (e.g., the respondent possesses more depression). The shape of this growth follows a monotonically increasing S-shaped curve called the item characteristic curve (ICC) which is defined by the item parameter(s).

The one parameter logistic model, also known as the Rasch model (Rasch, 1960), is a common IRT model. The probability of endorsing an item for a given level of θ is modeled as

$$P_i(\theta) = \frac{1}{1 + \exp[-(\theta - b_i)]} \ , \tag{7}$$

where b_i is the item difficulty parameter. For this model, the difficulty parameter indicates the point on the latent continuum where the probability of endorsing the item is .5.

The two parameter logistic (2PL) model is another popular item response model (Birnbaum, 1968). The 2PL includes two item parameters, difficulty and discrimination. The probability of endorsing an item for a given level of θ is modeled as

$$P_i(\theta) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]} \ , \tag{8}$$

where b_i is the item difficulty parameter, and a_i is the item discrimination parameter for this model. The discrimination parameter describes how well the item differentiates between respondents above and below the point at *b*.

Birnbaum's (1968) three-parameter model (3PL) is a common IRT model that adds an item parameter to account for the impact of guessing on item response. The probability of endorsing an item for a given level of θ is modeled as

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp[-a_i(\theta - b_i)]}, \tag{9}$$

where c_i is the pseudo-guessing parameter, b_i is the difficulty parameter, and a_i is the discrimination parameter.

Previous researchers have demonstrated the connections between IRT- and SEM-based (i.e., factor analytic) analyses of categorical measurement models (e.g., Takane & de Leeuw, 1987). If dichotomous or categorical data are used, factor analytic and IRT methods may be employed to estimate equivalent models (e.g., Edwards & Wirth, 2009). Specifically, the factor loading parameter can be transformed into a discrimination parameter, and the item intercept can be transformed into an item difficulty parameter (Kamata & Bauer, 2008). Although IRT and SEM have traditionally employed a separate vocabulary and different computer software programs, a growing literature relating IRT and factor analysis has expanded our understanding of the connections between the two frameworks (e.g., Takane & de Leeuw, 1987; Kamata & Bauer, 2008; Edwards & Wirth, 2009; Reise, Widaman, & Pugh, 1993; Chan, 2000).

Linking Under IRT and CTT

Holland (2007) provides a framework for organizing methods for linking scores from one test to another using CTT and IRT methods. He defines *linking* as a transformation of a score on one test to a score on another test. In this section, I describe types of linking, data collection designs related to linking, and options related to the technical process of linking using CTT and IRT methods, and I compare these to the SEM-based approaches to linking embodied by the shifting indicators model.

Types of linking. Two major sub-categories within Holland's (2007) typology are scale aligning (also called scaling), and equating. *Scale aligning* involves making scores on different tests comparable. The types of tests to be linked may measure different or similar constructs, may have similar or dissimilar psychometric properties (i.e., difficulty, reliability), and may be taken by a common or different populations of examinees. *Equating* is the process of making scores on different tests interchangeable. Equating requires more restrictive assumptions than scale aligning. Specifically, the tests to be linked must measure the same construct with comparable levels of reliability and difficulty, which implies that the tests should be built to the same specifications and administered under comparable conditions (Holland, 2007).

The shifting indicators design shares some commonalities with both scale aligning and equating. For example, there are similarities between the shifting indicators model and calibration, which falls under scale aligning in Holland's (2007) typology. According to Holland, calibration involves linking tests that measure similar constructs with different reliability taken by a common population. A common use of calibration involves linking scores from a short form of a longer assessment. The shifting indicators model may also involve the use of a truncated assessment (i.e., an assessment where items have been dropped), although the assessment could be the same length over time if dropped items are replaced with other items. Within the scale aligning category, the shifting indicators model also bears resemblance to vertical scaling. A common application of vertical scaling is to put achievement tests taken by students in adjacent grades on the same scale. Under vertical scaling, the populations of test takers are different (e.g., third graders and fourth graders), while under the shifting indicators model, the population of test takers is the same (e.g., the same population assessed as third graders one year and fourth graders the next year).

The shifting indicators model also shares features of equating. Holland (2007) differentiates equating from other forms of linking by stating that "[w]hat distinguishes test equating from other forms of linking is its demanding goal of allowing scores from both tests to be used interchangeably for any purpose" (p. 22). Like equating, the shifting indicators model imposes restrictive assumptions on the data (such as the equality constraints on loadings and intercepts over time). The goal of the shifting indicators model is to obtain measures of the latent construct that can be considered interchangeable in the sense that the scores are used to interpret growth. If scores are not reflective of the amount of the latent construct possessed by the respondents at each time point, interpretations of growth in that latent construct will be flawed.

Data collection designs. Many different data collection designs for scaling and equating tests have been developed. A major distinction among these methods pertains to the use of a common population or a common set of items.

Among the methods that utilize a common population, a *single group design* involves having a single group of respondents complete both forms of a test. The *counterbalanced design* is similar to the single group design, but varies the order of the two tests among respondents to counteract fatigue and other order effects. The *equivalent group design* (also known as the *random groups design*; Kolen & Brennan, 2004, pp. 13-15) entails randomly assigning members of a common population to take different forms of a test.

Another category of equating procedures relies on a common set of items rather than a common population. This type of design has been called the *non-equivalent groups with anchor test design*, also known as NEAT (Holland, 2007). Under the NEAT design, a common set of items, known as an anchor test, is used to link different tests taken by different populations of respondents. The anchor test is designed to reflect the content and statistical properties of the larger test.

The shifting indicators design has some commonalities with equating procedures that call on common items and those that call on a common population. As described earlier, the shifting indicators model accomplishes linking based on the presence of common items. However, the shifting indicators design also involves a common population of respondents over time. A key difference between the data collection designs described here and the shifting indicators model is that the shifting indicators model is a longitudinal model. Longitudinal IRT models are possible and are introduced in a subsequent section.

Technical processes of linking. The technical process of linking may be accomplished given a number of different methods. Mean equating, linear equating, and equipercentile equating employ a classical test theory framework. *Mean equating* adjusts scores on different forms of a test so that they have the same mean (Kolen & Brennan, 2004, p. 30). *Linear equating* adjusts scores to give the two forms of a test the same mean and standard deviation (Kolen & Brennan, 2004, p. 31). *Equipercentile equating* adjusts scores to give the two forms of the test approximately the same score distribution (Kolen & Brennan, 2004, pp. 36-37).

IRT methods may also be used for linking. For the data collection designs that involve non-equivalent groups and common items, three types of available procedures are concurrent calibration, fixed common item parameter calibration, and separate calibration.

Under *concurrent calibration*, all parameters are estimated simultaneously given the full set of data available (Lord, 1980, p. 201). That is, responses on both forms of the test are collapsed into a single dataset, and item and ability parameters are estimated together in one run. Responses are coded missing for the items on the test form that each set of examinees did not take. The item parameters of the anchor items are assumed to be the same (i.e., invariant) across the two forms of the test; thus, the metric of test forms is set by these common items. *Fixed common item parameter calibration* is similar to concurrent calibration, however, the parameters of the common items are already known (e.g., are pulled from an existing item bank) and are fixed across the different forms *a priori* (Jodoin, Keller, & Swaminathan, 2003).

Separate calibration involves estimating item parameters for each test form separately and then linking the forms using items that are common to both forms. Under separate calibration, the goal is to obtain transformation constants, typically called the slope constant and the location constant (or A and B; Kolen & Brennan, 2004, p. 163) that place the parameters from different forms on the same scale.

One way to obtain these transformation constants is via the item parameter estimates of the common items. The *mean/sigma method* uses the means and standard deviations of the *b*-parameters to calculate the transformation constants (Kolen & Brennan, 2004, p. 167). After transformation, the mean and standard deviation of the *b*-parameters will be identical for the two forms of the test. The *mean/mean method* uses the means of the *a*-parameters and the *b*-parameters to calculate the transformation constants (Kolen & Brennan, 2004, p. 167). The outcome of the mean/mean method is that the means of the *a*- and *b*- parameters will be the same for both forms of the test.

Characteristic curve methods are also suitable for designs involving non-equivalent groups and common items. These methods consider all of the item parameters simultaneously by comparing the difference between the item characteristic curves of the common items taken by different populations. The differences in test characteristic curves may also be considered. The goal of these methods is to obtain the transformation constants that minimize this difference. Two common characteristic curve methods are the Haebara approach and the Stocking and Lord approach (Kolen & Brennan, 2004, pp. 168-169).

In comparison to the CTT and IRT methods described above, the shifting indicators model employs a linking process that is most similar to concurrent calibration. The shifting indicators model estimates item and ability parameters in a single run, and sets the metric via equality constraints placed on the common items. That is, for the set of common items, the factor loadings and item intercepts are set to be equal across time points. In a case where item parameters are known ahead of time, it would also be possible to fix them *a priori* in a manner similar to the fixed common item parameter calibration method.

Longitudinal IRT Methods

Another important aspect of linking in the shifting indicators model pertains to the longitudinal nature of the data. Many of the CTT and IRT methods described to this point are generally employed in cross-sectional contexts, such as large-scale achievement testing. However, IRT methods are being used in an increasingly wide variety of measurement contexts, including longitudinal settings. In this section, I describe some of the methods for longitudinal item response theory (LIRT) and contrast them with the shifting indicators model.

Many of the LIRT models operate under the Rasch framework. The linear logistic test model (LLTM: Fischer, 1973) is an example of an item response model that was later extended to the longitudinal case (Fischer, 1989). A longitudinal extension of this model is the linear logistic test model with relaxed assumptions (LLRA). At the first time point, the probability of item response may be expressed as

$$P(x_{ij} = 1 \mid \theta_{ij}) = \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}, \tag{10}$$

where θ_{ij} is an item-specific ability parameter. After the initial time point, other effects are added to the model such that

$$P(x_{ij} = 1 \mid \theta_{ij}) = \frac{\exp(\theta_{ij} + \delta_{ij})}{1 + \exp(\theta_{ij} + \delta_{ij})}, \tag{11}$$

where δ_{ij} is the sum of the changes undergone by the respondent between time 1 and time 2. This change parameter typically includes a group-specific treatment effect and a trend effect capturing change not related to the treatment.

Several distinctive features characterize this model. A stepwise model testing procedure is employed to draw conclusions about growth and treatment effects; the fit of a baseline model is compared to more restrictive models that constrain the change parameter(s) to be equal across time and/or treatment group, in order to draw conclusions about growth. Drawbacks to the model are that (a) it measures only group-level change, (b) any change in item parameters is assumed to be the same for all items, and (c) estimation can be problematic when there is change in predominately one direction, such as measuring children's motor skills over time (Glück & Spiel, 1997).

Andersen (1985) proposed another Rasch-based model for estimating latent abilities of examinees at different time points. Andersen's model may be expressed as

$$P(x_{i(k)j} = 1 \mid \theta^*_{jk}, b_i) = \frac{\exp(\theta^*_{jk} - b_i)}{1 + \exp(\theta^*_{jk} - b_i)}, \tag{12}$$

where θ^*_{jk} is the ability of person *j* at time *k*, and b_i is the difficulty for item *i*. One defining feature of this model is that latent abilities are occasion-specific. These abilities are assumed to be different (although correlated) at each time point. The difficulty parameter for each item is assumed to be constant over time. As with all Rasch models, the item discrimination parameters are set to equality for all items at each time point. Thus, the model assumes longitudinal measurement invariance at the metric and scalar levels.

Andrade and Tavares (2005) proposed a 3PL longitudinal IRT model, which von Davier, Xu, and Carstensen (2011) classified as an extension of the Andersen (1985) model. The model may be expressed as

$$P(x_{i(k)j} = 1 \mid \theta^*_{jk}, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[a_i(\theta^*_{jk} - b_i)]}{1 + \exp[a_i(\theta^*_{jk} - b_i)]}, \tag{13}$$

where θ^*_{jk} and b_i are as previously defined, and a_i and c_i are the discrimination and pseudo-guessing parameters, respectively. In this model, the item parameters are assumed to be known and fixed over time, and ability parameters are estimated at each time point. Therefore, this model also assumes longitudinal measurement invariance at the metric and scalar levels.

Another model for longitudinal measurement is the multidimensional Rasch model for learning and change (MRMLC: Embretson, 1991). Embretson's model may be expressed as

$$P(x_{i(k)j} = 1 \mid \theta_j, b_i) = \frac{\exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}{1 + \exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}, \tag{14}$$

where θ_j is a vector of examinee abilities where θ_{j1} is the initial ability at the first time point (i.e., k = 1) and θ_{j2} through θ_{jm} are modifiabilities, and b_i is the difficulty for item *i*. A distinguishing characteristic of the MRMLC is the inclusion of "modifiabilities" that represent individuals' gain or decline in ability over time. A simplex structure links the probability of item response to initial ability and one or more modifiabilities. In this model, different items are administered at each time point. Items are paired on difficulty, so item 1 at time 1 has the same difficulty as item 1 at time 2, for example. Therefore, longitudinal measurement invariance at the metric and scalar levels is assumed, even though different items are used across time.

Another approach to LIRT involves incorporating IRT parameters into a hierarchical linear modeling (HLM) framework. This hybrid of the Rasch model and multilevel modeling is known as the hierarchical generalized linear model (HGLM: Kamata, 2001). Pastor and Beretvas (2006) demonstrated an extension of this model to the longitudinal context. A total of three levels are modeled: items at level 1, time at level 2, and persons at level 3. The first level models the log odds of endorsing an item as a function of a) an overall effect (i.e., common to all items) related to the examinee's latent ability at time *k* and b) an item effect related to the item difficulty at time *k*. The second level models variation in latent trait estimates and item effect estimates within persons over time. The third level models variations in growth among persons over time. An advantage of this approach is the ability to draw conclusions at both the individual and the item level. That is, this approach allows for the investigation of changes in latent trait for individuals and changes to the item parameters over time. (However, longitudinal measurement invariance can only be examined at the scalar, and not the metric, level because the model is Rasch-based.)

Similarities between HLM and latent growth modeling (LGM) have been described in detail (e.g., Preacher, Wichman, MacCallum, & Briggs, 2008). HLM and LGM models may be specified to obtain equivalent key parameters (i.e., slope mean and variance, intercept mean and variance, slope-intercept covariance). Given this connection—along with the relationship between IRT discrimination and SEM intercept parameters—it seems possible to specify equivalent HGLM and second-order LGM models, although I am not aware of any published works expositing this link.

While IRT- and SEM-based methods may both be used to develop a common metric for measures that change over time, second-order LGMs with shifting indicators offer several advantages that may make them attractive to researchers. First, not all applied researchers have training in IRT methods (especially longitudinal IRT) or access to IRT software. Researchers may be more inclined to use analytical techniques that are familiar to them, and second-order LGMs with shifting indicators combine two popular techniques: confirmatory factor analysis and latent growth curve modeling. Further, IRT software may not be equipped to estimate growth models. Second-order LGMs with shifting indicators estimate the measurement model and the growth model simultaneously. This process may be less demanding than estimating these parameters in separate steps and possibly different software programs. Third, it appears that longitudinal SEM methods are currently more flexible than longitudinal IRT methods with regard to testing for longitudinal measurement invariance and allowing for partial measurement invariance if necessary. Of the LIRT methods described above, only the longitudinal extension to the HGLM (Pastor & Beretvas, 2006) allows item parameter estimates to vary over time.

However, because this method is based on the Rasch model, longitudinal measurement invariance can only be investigated at the scalar level. As psychological and behavioral researchers look to adopt more sophisticated methods for examining longitudinal growth, second-order LGMs with shifting indicators may represent an accessible and user-friendly method of investigating longitudinal trajectories given a set of items that change over time.

Conditions Affecting the Accuracy of Second-order LGMs with Shifting Indicators

To date, empirical investigations of the performance of second-order LGMs with shifting indicators have been quite limited in scope. Hancock and Beuhl (2008) describe the model and demonstrate its use, but they do not evaluate the performance of the model under various manipulated conditions (e.g., by varying the number of items, loading magnitudes, etc.). Although research on the conditions affecting the accuracy of second-order LGMs with shifting indicators is limited, the literature related to models with similar features, including IRT models, CFA models, and other LGMs, may be useful in drawing inferences about which conditions may impact the performance of the shifting indicators model. In particular, I consider connections from related models to the impact of the proportion of shifting indicators in the model, the loading magnitudes of shifting indicators, and sample size.

Proportion of Shifting Indicators in the Model

One area of interest that has not been investigated in second-order LGMs with shifting indicators relates to the proportion of shifting indicators in the model. Two relevant conditions in this area are (a) the proportion of shifting to non-shifting indicators within a measurement occasion (e.g., two out of eight items are dropped versus four out of eight) and (b) the proportion of measurement occasions that involve shifting indicators (e.g., items are dropped at one time point versus three time points). In other words, does the performance of the shifting indicator model vary given a small versus large proportion of shifting indicators within one measurement occasion? And, does the performance vary given a small versus large proportion of measurement occasions with shifting indicators? A related issue pertains to how the number of indicators per factor impacts the quality of measurement because a larger number of shifting indicators results in a smaller number of items per factor unless new items are added to replace the dropped ones.

Work in equating and vertical scaling indicates that employing a larger set of common items is preferable (e.g., Kolen & Brennan, 2004, p. 271, Fitzpatrick, 2008). A larger set of common items has been associated with a reduction in equating error (Budescu, 1985; Wingersky, Cook, & Eignor, 1987). One rule of thumb that has been suggested is that the set of common items should comprise at least 20% of the total test length (Cook & Eignor, 1991).

In SEM models, researchers have investigated the relationship between the number of indicators per factor and the quality of the measurement model. Using a series of simulation studies, Marsh, Hau, Balla, and Grayson (1998) demonstrated that using more indicators per factor resulted in greater model convergence, more accurate parameter estimates, more stable parameter estimates, and greater reliability of factors. For model convergence, the advantage of having a larger number of indicators per factors was amplified in cases where sample size was small. In one example, given a sample size of 50, a proper solution was obtained 99% of the time with nine indicators per factor and only 41% of the time with three indicators per factor. For parameter estimates, increases in the number of items per factor were related to more accurate estimates of factor loadings, uniquenesses, and factor correlations, although this effect appeared to level out once the number of items per factor reached four. However, the standard deviations of those estimates continued to systematically decrease as the number of items per factor increased. Additional items per factor were also related to increased factor reliability. The relationship between factor reliability and model estimation is described in the next section.

Building on the work of Marsh, Hau, Balla, and Grayson (1998), Gagné and Hancock (2006) reported advantages of using more versus fewer indicators per factor in a related study. They found that increasing the number of indicators per factor generally leads to improvement in model convergence and parameter estimation in CFA models. They also found an interaction between the number of items per factor and sample size on model convergence; these facets of the model worked in a compensatory fashion, with larger values in one area able to make up for smaller values in the other. For example, with four indicators per factor, an average sample size of 1,000 was required to meet a satisfactory rate of convergence (i.e., ≤ 1,100 samples generated to attain 1,000 properly converged replications) across tested conditions, whereas with 12 indicators per factor, the average necessary sample size to meet this criterion was 200. In terms of the accuracy of parameter estimates, they found that under conditions that had a relatively high likelihood of model convergence (i.e., those that met the "satisfactory" criterion described above), parameters, standard errors, and chi-square estimates were generally accurate, although standard errors and chi-square values were occasionally inflated. The authors did not report results of these bias analyses for each of the independent variables (i.e., number of items per factor, sample size, loading magnitudes) separately; however, they concluded that more indicators per factor, larger sample size, and larger loading magnitudes increase the likelihood that the model will converge and provide unbiased estimates.

A study investigating the performance of second-order LGMs indicated that a greater numbers of items per factor may have an adverse impact on bias in estimates of the chi-square statistic given small sample sizes. Leite (2007) found that the relative bias of the chi-square statistic increased as the number of items per factor increased from five to 10 and from 10 to 15. However, at sample sizes of at least 500, the level of bias did not increase beyond an acceptable level. Similar results were obtained for the CFI and the Tucker-Lewis index (TLI).

This inverse relationship between fit as assessed by the chi-square fit statistic and the number of items per factor was also observed by Marsh, Hau, Balla, and Grayson (1998). They found systematic variation among properly specified CFA models with regard to the number of indicators per factor, with goodness of fit appearing to be worse as items per factor increased, especially at smaller sample sizes.

Given these findings, it is expected that models that employ a greater proportion of non-shifting items (i.e., have more common indicators per factor within each measurement occasion and across time) will result in better rates of proper convergence, more accurate parameter estimates, and more stable parameter estimates in comparison to models with fewer common items. However, models with a greater number of items per factor (i.e., fewer shifting items) may exhibit larger values of relative bias with regard to the chi square (and related) fit statistics, especially when sample size is small.

Loading Magnitudes of the Shifting Indicators

Researchers interested in second-order LGMs with shifting indicators may also want to better understand how the magnitude of the loadings of the omitted items relates to the performance of the model. Items that are identified based on theory as being developmentally inappropriate at certain time points are likely to be poor indicators of the latent construct at those measurement occasions. Therefore, if they are included in the model, they likely will have low factor loadings and will not be contributing much reliable variance at those times. On the other hand, in some cases indicators with high loadings may be omitted from some measurement occasions due to unexpected occurrences--such as problems with data collection. While no other studies have investigated how the loading magnitudes of omitted items relates to the performance of second-order LGMs with shifting indicators, connections can be drawn from research in confirmatory factor models.

Gagné and Hancock (2006) evaluated model convergence and parameter estimation of confirmatory factor models given varying levels of loading magnitude, sample size, and number of indicators per factor. The results of this study were briefly described in the previous section and are elaborated on here. The authors found that estimation improved as levels of the independent variables increased; higher loadings, larger samples, and more indicators per factor resulted in better convergence rates and more accurate parameter estimates and variances. With regard to factor loadings, they demonstrated that models with higher factor loadings required a smaller sample size to obtain satisfactory rates of model convergence. A satisfactory level of convergence was defined as less than or equal to 1,100 samples generated to attain 1,000 properly converged replications. For example, given six indicators per factor with loading magnitudes of .8, a sample size of 25 was needed to meet the criterion for satisfactory convergence. In a 6-indicator model with loadings of .2, a sample size of 1,000 was needed.

One contribution of this study that is particularly relevant to the shifting indicators model is the investigation of the impact of various configurations of heterogeneous loading magnitudes on estimation of CFA models. While Marsh, Hau, Balla, and Grayson (1998) only investigated models with loading magnitudes of .6 for all items, Gagné and Hancock (2006) investigated models given four different levels of loading magnitudes: .2, .4, .6, and .8. Further, they generated models with homogenous loadings (i.e., all loadings are the same for each indicator in the model), heterogeneous loadings (i.e., loadings differ within the model), and with different numbers of indicators. This arrangement allowed the researchers to investigate questions such as whether adding additional items per factor improves estimation when the loadings of the added items are low. This line of inquiry may be particularly pertinent to the shifting indicators model given that researchers may choose to omit items with low loadings at selected time points.

The results of the study indicated that in almost all cases, adding indicators to the model improved convergence, including cases where the added indicators had loadings of .2. The only exception to this pattern was in the case where the sample size was less than 50; in this case, the addition of multiple indicators with loadings of .2 decreased the convergence rate under some conditions.

In summarizing the impact of loading magnitude and number of indicators per factor on model estimation, the authors focused on construct reliability. Construct reliability is a function of the number of items per indicator and the factor loadings of those items and is often estimated using coefficient omega, expressed as

$$\omega = \frac{\left(\sum_{i=1}^{k} a_i\right)^2}{\left(\sum_{i=1}^{k} a_i\right)^2 + \sum_{i=1}^{k}(1-a_i^2)}, \tag{15}$$

where a_i refers to the loadings for the set of *k* indicators (Gagné & Hancock , 2006). The authors conclude that as construct reliability is enhanced (i.e., the number of indicators per factor and/or loading magnitudes increase), the likelihood that the model will converge will also increase.

A drawback to this study is that the impact of loading magnitudes or any of the other studied conditions on bias in other quantities (i.e., parameter and standard error estimates, and chi-square values) was not reported separately. However, the authors in general linked the likelihood of convergence to reductions in bias. They concluded that more indicators per factor, larger sample size, and larger loading magnitudes increase the likelihood that the model will converge and provide accurate estimates.

The results of this study provide several implications for the shifting indicators model. First, in terms of convergence, it may be undesirable to voluntarily drop items from a study, even when the loadings of those items are low. If the items are available, it may be preferable to include them. However, omitting items with low loadings from the shifting indicators model should have less of an impact on the estimation of the model than omitting items with higher loadings. Researchers can determine the loss of construct reliability caused by omitting an item with a given loading magnitude using the formula for coefficient omega.

Sample Size

Applied researchers may be particularly interested in recommendations related to adequate sample size in the shifting indicators model. Based on work in confirmatory factor analysis (CFA) and latent growth curve modeling contexts, it is expected that the quality and stability of estimation of shifting indicators models will improve as sample size increases. However, a key question relates to the minimum sample size needed in order for researchers to expect reasonably stable and accurate results given a second-order LGM with shifting indicators. The answer to this question will depend on the characteristics of the data and the model.

Leite (2007) demonstrated that second-order LGMs with identical indicators performed well under a variety of conditions with sample sizes of at least 500. The independent variables of this simulation study were sample size, number of items, number of time points, equal versus unequal loadings across items, level of reliability, and level of longitudinal measurement invariance. The dependent variables were percentage of inadmissible solutions, bias of the parameter estimates and standard errors, bias of the chi-square statistic, and performance of other fit indices (CFI, TLI, RMSEA).

In terms of admissible solutions, sample size and the number of measurement occasions impacted the rate of convergence to a proper solution. Given a sample size of 100, the percentage of inadmissible solutions was 32.4% with three measurement occasions and 0.9% with five measurement occasions. With a sample size of 1000, 8.9% of solutions were inadmissible with three measurement occasions, while no inadmissible solutions were produced given five measurement occasions. This result indicates that sample size and number of measurement occasions can serve a compensatory function for each other.

With regard to bias, second-order LGMs were shown to accurately estimate the key growth parameters. Under all conditions except one, the relative bias of the growth parameters and standard errors were within the acceptable limit (i.e., .05 for parameter estimates and .1 for standard errors; Hoogland & Boomsma, 1998). The anomalous result occurred in the condition with a sample of 100, 15 items, and five measurement occasions. In this case, the relative bias of one of the parameter estimates (i.e., the slope/intercept covariance) exceeded the acceptable limit. Overall, this result suggests that even when sample size is small, the accuracy of estimating growth parameters is suitable given second-order LGMs.

Sample size played a more influential role in terms of the performance of fit indices. In the conditions where sample size was 100 or 200, the relative bias of the chi-square statistic exceeded the acceptable limit (i.e., .05). Furthermore, CFI, TLI, and RMSEA fit indices demonstrated similar results. When the sample size was 500 or 1000, the percentage of analyses indicating adequate fit according to the CFI, the TLI and the RMSEA was at or near 100%. With sample sizes of 100 or 200, the pattern was more complicated. When the number of measurement occasions and the number of items per factor was smaller, and the construct reliability of the factors was higher, a higher percentage of fit indices indicated adequate fit. For example, with a sample size of 200, three time points, and five items per factor, and a construct reliability of .9, the adequate fit percentage exceeded 98% for the CFI, the TLI, and the RMSEA. (Note that the RMSEA was not impacted by construct reliability.) On the other hand, in conditions with more time points, more items per factor, and lower construct reliability, the performance of the fit indices declined. For example, with five time points, 15 items per time point, and a sample size of 100, the percentage of the analyses with fit indices indicating adequate fit was zero, regardless of the construct reliability.

Based on the problematic performance of the chi-square statistic and the other fit indices investigated under small sample sizes, Leite (2007) recommends the use of the second-order LGMs only when the sample exceeds 500. In an applied study, Pettit, Keiley, Laird, Bates, and Dodge (2007) successfully employed a shifting indicators model given a sample size of 522.

Given the manipulated variables in this study, it is expected that a relatively larger sample size will be required when the proportion of shifting indicators is larger. This study aims to further clarify the relationship of sample size to the other independent variables to help researchers decide whether second-order latent growth modeling with shifting indicators may be appropriate given the characteristics of their data.

Research Questions and Rationale

Initial evaluations of the performance of second-order LGMs with shifting indicators have been promising. Due to these positive results, additional work to determine the performance of second-order LGMs with shifting indicators under a wider range of conditions is warranted. The goal of this study is to provide a clearer understanding of the performance of second-order LGMs with shifting indicators under a variety of conditions so that researchers are better informed about when and how the use of the model may be appropriate. Simulated data were generated with all items present at each time point. Given a common generating data set, the performance of two types of LGMs were compared: (1) a second-order LGM with shifting indicators, and (2) a second-order LGM with all items present at each time point (i.e., a “full model”).

The shifting indicators model is of primary interest in this study. This model represents the case where items have been omitted from the measure on theoretical or practical grounds. The full model served as a comparison model. This model represents the case where all items are available and included in each time point.

Monte Carlo simulation was employed to address four research questions. For each question, the performance of a second-order LGM with shifting indicators was compared to the full LGM. The research questions and associated hypothesis are:

1. How does the number of shifting to non-shifting indicators within the affected measurement occasions influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?
   It is hypothesized that models with more indicators per factor within each measurement occasion will have better rates of proper convergence, less bias in parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation. With regard to fit, models with more indicators per factor within each measurement occasion are hypothesized to have larger model rejection rates based on the chi-square (i.e., Type I errors) and related fit statistics.

2. How does the number of measurement occasions with shifting indicators influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?

   It is hypothesized that models with fewer occasions involving shifting indicators will have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation. With regard to fit, models with fewer occasions of shifting indicators are hypothesized to have larger model rejection rates based on the chi-square (i.e., Type I errors) and related fit statistics.

3. How does the magnitude of the factor loadings for the omitted items influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?

   It is hypothesized that models that drop items with low loadings will have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation than models that drop items with high loadings. With regard to fit, models that drop items with low loadings are hypothesized to have smaller model rejection rates based on the chi-square (i.e., Type I errors) and related fit statistics.

4. How does sample size influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?

   It is hypothesized that models with larger sample sizes will have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation than models with smaller sample sizes. With regard to fit, models with larger sample sizes are hypothesized to have smaller model rejection rates based on the chi-square (i.e., Type I errors) and related fit statistics than models with smaller sample sizes.

The research questions addressed in this study are intended to focus on areas that would be of practical interest to applied researchers faced with making a decision about whether to use the shifting indicators model given the characteristics of their data. In the next chapter, I elaborate on the operationalization of these research questions.

CHAPTER 3
METHODS

This study evaluated the performance of second-order LGMs with shifting indicators under several different conditions. Specifically, the number of shifting to non-shifting indicators, the number of occasions with shifting indicators, the magnitude of the factor loadings of the shifting indicators, and sample size were examined. These were manipulated to investigate the impact on model convergence, growth parameter and standard error estimation, efficiency, and model fit. Monte Carlo simulation was employed to generate data where all indicators are present at all time points. Data for selected indicators at selected time points were then deleted to mimic a design in which shifting indicators are encountered. The performance of the second-order LGM with shifting indicators was then compared to the model with all indicators present at all time points.

Independent Variables

Five independent variables were manipulated in this study. They are:

1. model type (two levels: full and shifting);
2. the magnitude of the factor loadings of the shifting indicators (two levels);
3. the number of shifting indicators per measurement occasion (three levels);
4. the number of measurement occasions with shifting indicators (four levels); and,
5. sample size (four levels).

For the purpose of this study, shifting indicators are defined as items that are dropped from models in the shifting condition and corresponding items in the full models that have low factor loadings (i.e., representing the case when the item is not developmentally appropriate). Table 3.1 summarizes the levels of the independent variables compared in this study.

Table 3.1
Levels of the Independent Variables

Independent Variable	Level
Model type	Full Shifting
Number of shifting indicators per occasion	0 2 4
Number of measurement occasions with shifting indicators	0 1 2 3
Loadings of items omitted in shifting conditions	.3 .7
Sample size	250 500 750 1000

Two types of second-order LGMs were compared: a full model with all indicators present at each measurement occasion, and a shifting indicators model. The models that included the full set of indicators served as a comparison for the shifting indicator models. Figure 3.1 depicts the basic components of a full model. For the full models, eight items measured the latent construct at each time point. A total of eight items were chosen to compose the full model to represent a realistic number of items for a scale measuring psychological or behavioral constructs. For example, the scales listed in *Measuring Bullying, Victimization, Perpetration, and Bystander Experiences: A Compendium of Assessment Tools* generally range from five to 10 items (Hamburger, Basile, & Vivolo, 2011). Further, a full model of eight items allows for sets of items to be dropped from the model while maintaining a minimum of four items per factor, as suggested by Marsh, Hau, Balla, and Grayson (1998).

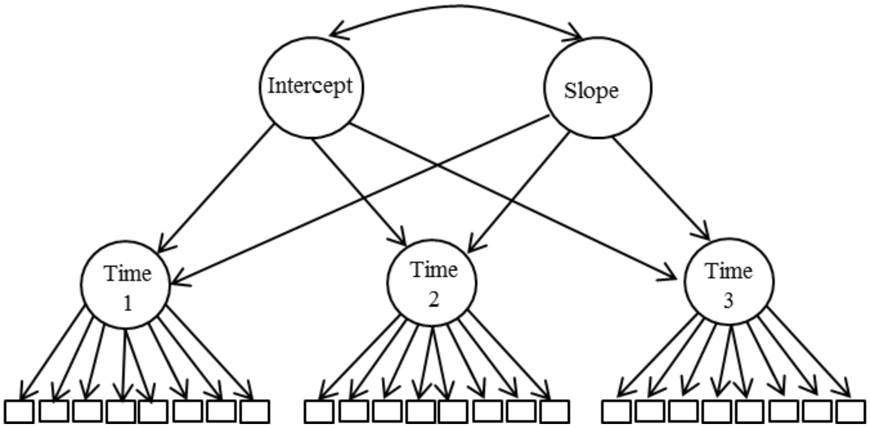


Figure 3.1. Basic components of a full model.

Three levels of the number of shifting indicators per measurement occasion were compared: zero, two, and four. When two items were dropped at the impacted measurement occasions, six items remained. This condition represents the case where a smaller number of items is shifted. When four items were dropped at the impacted measurement occasions, only four items remained. This condition represents a more extreme case where a larger number of items are shifted over time. Although this circumstance (i.e., half of all items are dropped) would not be ideal, it was of interest to investigate whether model performance would degrade when a large proportion of items were dropped.

Four levels of the number of measurement occasions with shifting indicators were investigated: zero times, one time, two times, and three times. In all cases, three measurement occasions were included in the model. In the one-time condition, items were dropped from the last occasion. This configuration is meant to reflect the case where some items are no longer relevant at the end of the study. In the two-times condition, items were excluded from the first and last measurement occasions. This configuration is meant to reflect a situation where some items are more relevant for participants at earlier stages of the study, and other items are more relevant for participants at later stages of the study. In the final condition, shifting indicators were included in all measurement occasions. Figure 3.2 depicts a simplified measurement model with two shifting indicators. Figure 3.3 depicts a simplified measurement model with four shifting indicators.

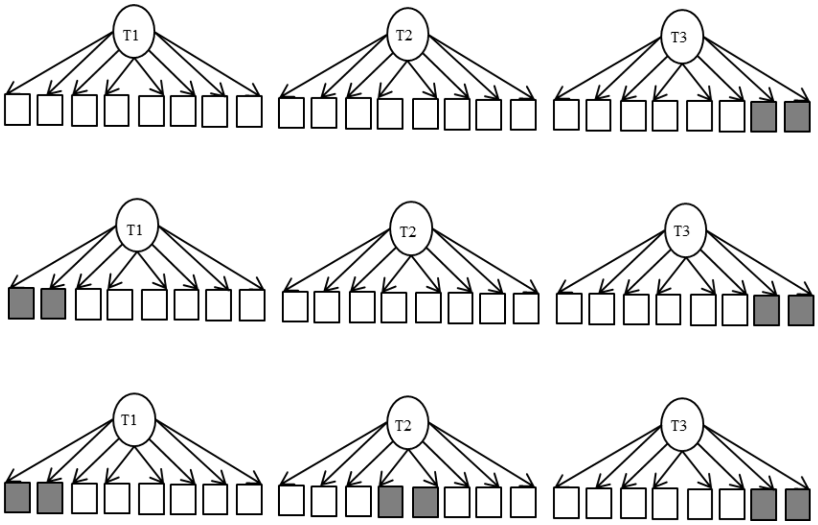


Figure 3.2. Simplified measurement model with two shifting indicators. Shifting items are designated with grey boxes. Top: One measurement occasion has shifting indicators. Middle: Two measurement occasions have shifting indicators. Bottom: All measurement occasions have shifting indicators. The pattern of shifting indicators has been configured to meet identification requirements.

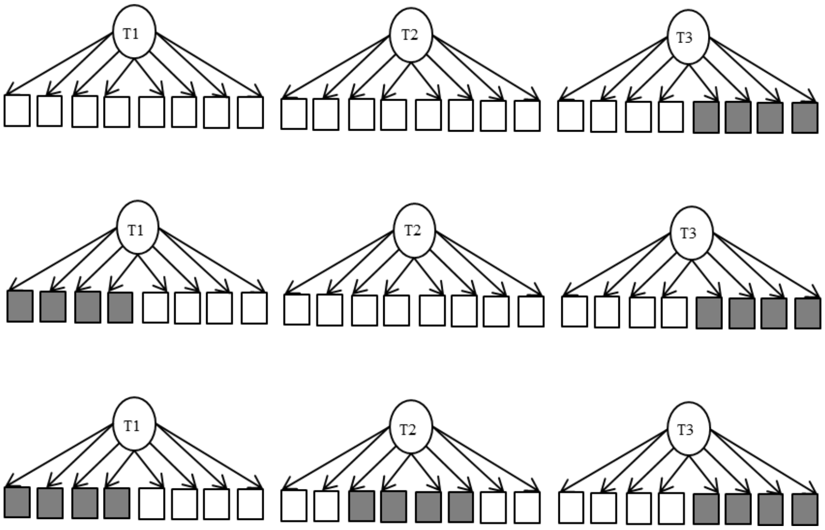


Figure 3.3. Simplified measurement model with four shifting indicators. Shifting items are designated with grey boxes. Top: One measurement occasion has shifting indicators. Middle: Two measurement occasions have shifting indicators. Bottom: All measurement occasions have shifting indicators. The pattern of shifting indicators has been configured to meet identification requirements.

Two levels of magnitude of the factor loadings for the items omitted in the shifting conditions were compared: .3, and .7. The loading magnitude of .3 is intended to reflect situations where items are less relevant at certain measurement occasions due to developmental changes. A loading magnitude of .3 is lower than the commonly used cut-point of .4 for selecting an item (Ding, Velicer, & Harlow, 1995). However, researchers may have reasons for wanting to keep a scale intact over time. For instance, when a scale has been previously validated and performs well (i.e., has high loadings) at most time points, researchers may wish to use the intact scale rather than drop items with low loadings. This level of loading magnitude was selected to investigate the impact of keeping items with low loadings in the model and to help determine whether the “more items are better” recommendation (e.g., Marsh, Hau, Balla, & Grayson, 1998; Gagné & Hancock, 2006) holds for second-order LGMs when loading magnitudes are low.

The second level of loading magnitude was .7. In this level, the loading magnitudes of the dropped items were the same as the intact items. This case is intended to reflect situations where relevant items are lost due to unplanned circumstances, such as problems with data collection. This level of loading magnitude was selected to investigate the impact of the inadvertent loss of items on model estimation to help determine the circumstances (if any) under which the performance of the model would be acceptable.

Four levels of sample size were compared: 250, 500, 750, and 1000. These values are similar in range to the larger values of sample size (i.e., >100) investigated by Leite (2007), Marsh, Hau, Balla, and Grayson (1998), and Gagné and Hancock (2006). The values were chosen at equally-spaced intervals to provide information along this continuum. Leite (2007) found that a minimum sample size of 500 was necessary for the adequate performance of full second-order LGMs under most conditions. For the purpose of this study, it was of interest to determine whether a similar degradation of performance is found with sample sizes of 250 given the different shifting model conditions that were manipulated in this study. The shifting indicator models have fewer parameters to estimate, so the sample size requirement might not be as large for the shifting models as the full models.

Dependent Variables

Five dependent variables were investigated in this study. They are:

1. model convergence;
2. bias in parameter estimates;
3. bias in standard error estimates,
4. efficiency, and;
5. model fit.

Rates of model convergence and inadmissible solutions were recorded for each of the conditions studied. When a model fails to converge or produces inadmissible solutions, it can present a major setback to the researchers' analytical plan. It is important for researchers to feel reasonably confident ahead of time that the model will converge and produce an admissible solution.

In the case when a model fails to converge, no usable output data (e.g., parameter estimates) is produced by the M*plus* program. In a simulation study, the number of models that fail to converge is apparent by the number of missing records at the end of the data analysis phase. For instance, when 998 output documents are produced by 1000 replications of a model, it can be deduced that two models failed to converge. For each combination of conditions, the number of models that failed to converge was recorded.

Improper solutions were also identified. Improper solutions are observed when the converged solution involves a non-positive definite variance/covariance matrix of the growth parameter estimates. In particular, negative values of the slope variance, intercept variance, or standard errors of the growth parameters were flagged. An overall convergence rate of at least 90% was considered satisfactory, in accord with Gagné and Hancock (2006).

The relative bias in estimates of the five key parameters of LGMs (i.e., slope mean, intercept mean, slope variance, intercept variance, slope/intercept covariance) was calculated. The relative bias in parameter estimates averaged across replications is calculated as:

$$Bias(\hat{\theta}) = \sum_{j-1}^{n_r} \left(\frac{\hat{\theta}_{ij} - \theta_i}{\theta_i} \right) / n_r, \tag{16}$$

where θ_i is the population parameter, $\hat{\theta}_{ij}$ is the parameter estimate for the *j*th sample, and n_r is the number of replications. When the relative bias of the parameter estimate is less than .05, it is considered acceptable (Hoogland & Boomsma, 1998). The relative bias estimates across replications were compared for each of the conditions using ANOVA. Within these five-way ANOVAs, the dependent variable is the relative bias of the parameter estimate, and the independent variables are model type (i.e., shifting vs. full), the number of shifting indicators, the number of occasions with shifting indicators, loading magnitude, and sample size. Effect size, measured by partial eta-squared, provides information regarding the level of practical significance of these results. Partial eta-squared (Cohen, 1973) is calculated as:

$$\eta^2 = \frac{SS_{between}}{SS_{between} + SS_{error}}. \tag{17}$$

An advantage of eta-squared is that the effects of other variables on the effect size estimate are controlled for, so results may be more easily compared across studies. In accordance with Leite (2007), partial eta-squared values of at least .05 were reported.

The relative bias in standard errors of the growth parameters (i.e., slope mean, intercept mean, slope variance, intercept variance, slope/intercept covariance) was calculated. The relative bias of the standard errors is calculated as:

$$Bias(S\hat{E}(\hat{\theta}_i)) = \sum_{j=1}^{n_r} \left(\frac{S\hat{E}(\hat{\theta}_i)_j - SE(\hat{\theta}_i)}{SE(\hat{\theta}_i)} \right) / \, n_r \ , \tag{18}$$

where $SE(\hat{\theta}_i)$ is the estimated population standard error of $\hat{\theta}_i$ and $S\hat{E}(\hat{\theta}_i)$ is an estimate of the standard error of $\hat{\theta}_i$ for the *j*th sample. A relative bias of the standard error of less than .1 is considered acceptable (Hoogland & Boomsma, 1998). The estimated population standard error, $SE(\hat{\theta}_i)$, can be calculated as the standard deviation of the parameter estimates across all replications. This value, as given in Bandalos (2006), can be calculated as:

$$\sqrt{\frac{\sum_{j=1}^{n_r}(\hat{\theta}_{ij}-\bar{\hat{\theta}}_i)^2}{n_r-1}}. \tag{19}$$

In this equation, $\bar{\hat{\theta}}_i$ is the average estimated parameter value across replications within a cell, and all other elements are as previously defined. Values of the relative bias of the standard errors were compared for each of the conditions using ANOVA, and partial eta-squared values of at least .05 are reported.

Efficiency is another criterion that is of interest in Monte Carlo studies. When comparing different methods of estimating the same parameters (in this case, the growth parameters of second-order latent growth models), the method that produces the lowest sampling variance is considered the most efficient. In this study, a simple measure of efficiency—the average standard error of the parameter estimate across replications, within each cell—is employed. Another common measure of efficiency is the mean square error (MSE; Paxton, Curran, Bollen, Kirby, Chen, 2001), which is the squared difference between the estimated parameter and the true population parameter. When parameter estimates are unbiased, MSE quantifies the sampling variability of the estimate. When bias exists in the parameter estimates, the MSE does not solely describe this variability but also incorporates bias (Enders, 2001). In this study, it was unknown *a priori* whether estimates produced by the shifting indicators model would be unbiased. Therefore, the average standard error of the parameter estimate was considered a more direct measure of efficiency.

Model fit was investigated using the chi-square fit index, the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). The chi-square fit index tests the null hypothesis that the population covariance matrix is equal to the model-implied covariance matrix. A discrepancy function, *F*, measures the difference between the matrices, and is minimized to produce $F_{\min}$. The test statistic $T = (N-1)F_{\min}$ has a large sample chi-square distribution, where $N-1$ is the model degrees of freedom. This value, commonly called the chi-square fit statistic, is compared against the critical value in the chi-square distribution at significance level α (Hu & Bentler, 1998). A non-significant chi-square value is desirable, as it indicates that the population and model-implied covariance matrices are not significantly different.

For every replication, the value of the chi-square fit statistic was compared to the critical value at the .05 level of significance; datasets that produced chi-square values that exceeded the critical value were considered to be rejected. Because the tested models fit the data perfectly in the population, rejected models represent Type I errors. The Type I error rate was calculated as the number of rejected models divided by the total number of replications within a cell. Rejection rates were compared across cells using Bradley's (1978) liberal criterion of $\alpha \pm .5\alpha$. In this study, $\alpha = .05$, so Type I error rates in the range of 2.5-7.5% were considered acceptable.

Although the chi-square fit index is a very popular measure of model fit, several drawbacks to its use have been reported (e.g., Bentler, 1990). For example, the chi-square index tests for *exact* fit between the population and model-implied matrices, when it is assumed that any given model will only produce an *approximation* of reality (Hu & Bentler, 1998). Also, the chi-square test is sensitive to trivial differences in the matrices under large sample sizes. Therefore, the chi-square test may be significant (indicating poor fit) even when differences between the population and model-implied covariance matrices are small. Because of these drawbacks, the chi-square fit statistic is often used in concert with other measures of model fit, such as the CFI, TLI, and RMSEA. Each of these other fit indices provides a descriptive measure of fit along a continuum, rather than a statistical test of model fit. The CFI, TLI, and

RMSEA were chosen to correspond with the fit indices investigated in Leite (2007) and because they are commonly used fit indices with established criteria for values indicative of good fit (Hu & Bentler, 1998, 1999).

CFI and TLI are incremental fit indices that compare the improvement in fit of the tested model versus a baseline model. The baseline model specifies that the observed variables have no relationship with one another. The TLI takes into account the parsimony of a model and is relatively insensitive to distribution and sample size. Values above .95 are recommended (Hu & Bentler, 1998, 1999). The CFI compares noncentrality parameters of the proposed and baseline models. It is also relatively insensitive to distribution and sample size. Values above .95 are considered indicative of good fit (Hu & Bentler, 1998, 1999).

The RMSEA is a stand alone fit index (i.e., it does not compare the tested model to a baseline) that is a measure of lack of fit. It corrects for model complexity and is insensitive to model type, sample size, and distribution. RMSEA values of less than .06 are desirable (Hu & Bentler, 1998), and a confidence interval is typically reported.

For each of these fit indices, the value of each index for every replication was compared to the criteria for acceptable fit described above. Each dataset was then classified as either meeting or not meeting the criteria for acceptable fit. Rejection rates for each of the fit indices were calculated as the number of rejected models divided by the total number of replications within a cell. Rejection rates were descriptively compared across cells.

Data Generation

Longitudinal data were generated using the external Monte Carlo function of the M*plus* software program (version 6; Muthén & Muthén, 2010). The SAS software program (version 9.3; SAS Institute, 2011) was used to interact with M*plus* and automate the data generation and analysis process.   Examples of SAS syntax to generate data and analyze full and shifting models are provided in Appendices A and B.

Samples were drawn from multivariate normal populations, and for each cell 1000 replications were conducted. This number of replications is consistent with Leite (2007). The population means for the intercept and slope were set to 1, in congruence with the parameter values assigned in Leite (2007). The intercept variance was set to 1, and the slope variance was set to .2. These values reflect the five to one ratio forwarded by Muthén and Muthén (2002) as being common in the applied literature. The slope/intercept covariance was set to .179, also in accordance with Leite. This value corresponds to a slope/intercept correlation of .4. The factor variances were set to 1, which in turn set the measurement error variances to one minus the squared loading of the item. Measurement error variances were uncorrelated. Factor loadings were set to .7 for the non-shifting items. Factor loadings in the range of .6-.8 are commonly seen in simulation studies (see, for example, the meta-analysis conducted by Hoogland & Boomsma, 1998). While some applied studies may have lower values of factor loadings, it was desirable to employ relatively high factor loadings in this study due to the number of dropped items within certain conditions. Because this is one of the first studies investigating the shifting indicators model, it is desirable to evaluate the performance of the model given high, but realistic, factor loadings for the non-shifting indicators. If the model performs well, future studies may evaluate the impact of lower overall factor loadings for the non-shifting indicators. Item intercepts were set to 0. If the results of this study are promising, future studies may evaluate the impact of non-zero item intercepts. An example of a generating model for one set of conditions, including population values for all parameters, is presented in Figure 3.4.

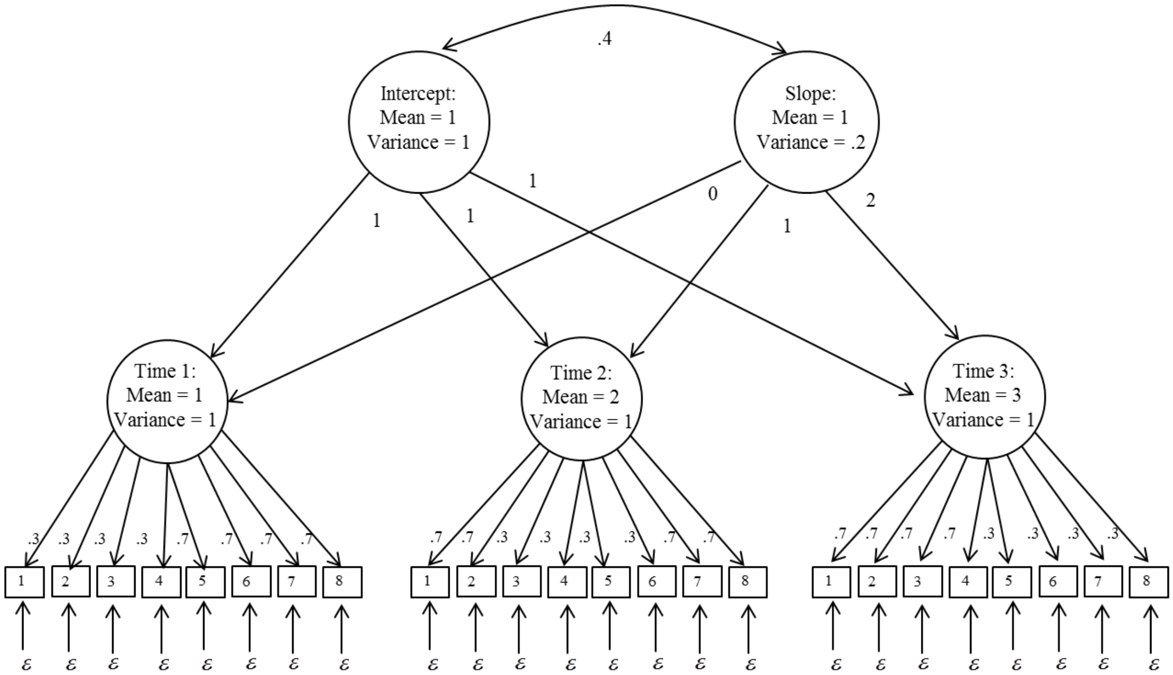


Figure 3.4. Example of a generating model with no indicators dropped. This example represents the full model for the set of conditions where four indicators are shifted at each of the three time points. The items to be shifted have a loading magnitude of .3 in the full model. Item intercepts are set to 0. Measurement error variances are set to one minus the squared item loading. Measurement errors are uncorrelated.

In total, 76,000 replications were conducted; that is, 1000 replications of 76 cells were obtained. The conditions of the study were not completely crossed. In the full model condition, simulated data were generated for 28 cells. After data were generated with all items present at each of these conditions, items were dropped according to the number of items and number of occasions conditions to create the shifting indicators models. There are 48 cells in the shifting model condition. Figures 3.5 and 3.6 provide a pictorial representation of the design of the study.

The discrepancy in the number of conditions under the full model and the shifting indicators model is due to the arrangement of the full models when the loading magnitude of shifting items condition takes a value of .7. It is easiest to describe this aspect of the study design by first elaborating on the development of the full models under the loading magnitude condition of .3. Under this condition, full models are created that include certain items with a loading magnitude of .3. The number and arrangement of the items with a loading of .3 depend on the levels of two other independent variables: the number of shifting items variable and the number of time points with shifting items variable. For example, when the number of items variable takes a value of two and the number of time points variable takes a value of three, the full model includes two items at every time point with a loading of .3, and all other items have a loading of .7. (The corresponding shifting indicator model drops the items with a loading of .3.) From this example, it can be seen that the full model will have a different combination of items with loadings of .3 and .7 for every level of the number of items and number of time points variables.

In contrast, this is not the case when the loading magnitude of items to be dropped condition takes a value of .7. In this case, in the full model all loadings are .7 regardless of the levels of the number of items and number of time points variables. Therefore, there is no difference in the full model regardless of the levels of these variables, as is reflected in Figure 3.4. However, the shifting model does change with each level of these variables, as different combinations of items are dropped. This feature of the design accounts for the difference in the number of cells under the shifting indicators model versus the full model.

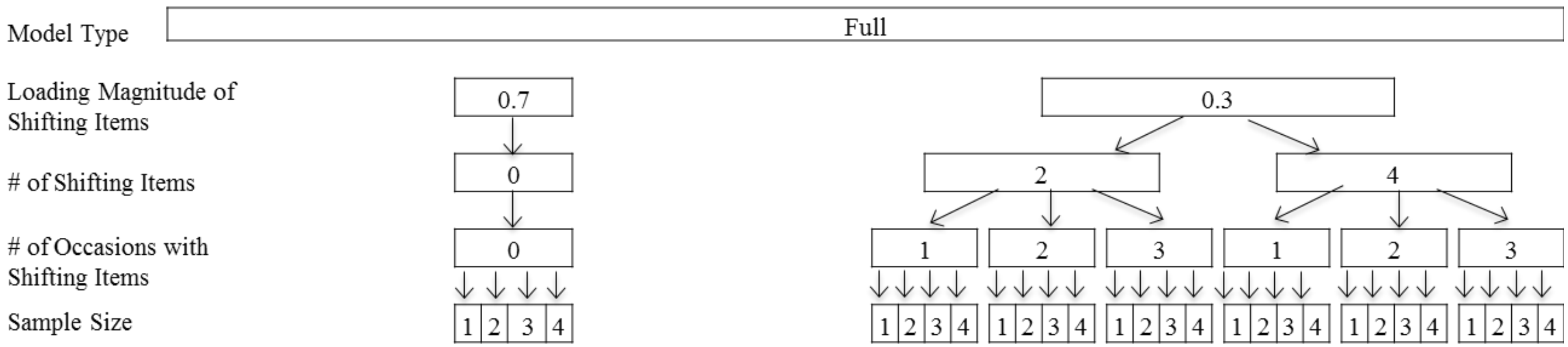


Figure 3.5. The 28 cells in the full model condition. The datasets generated under this condition serve as the complete datasets from which items are dropped for the shifting conditions. Sample size: 1 = 250, 2 = 500, 3 = 750, 4 = 1000.

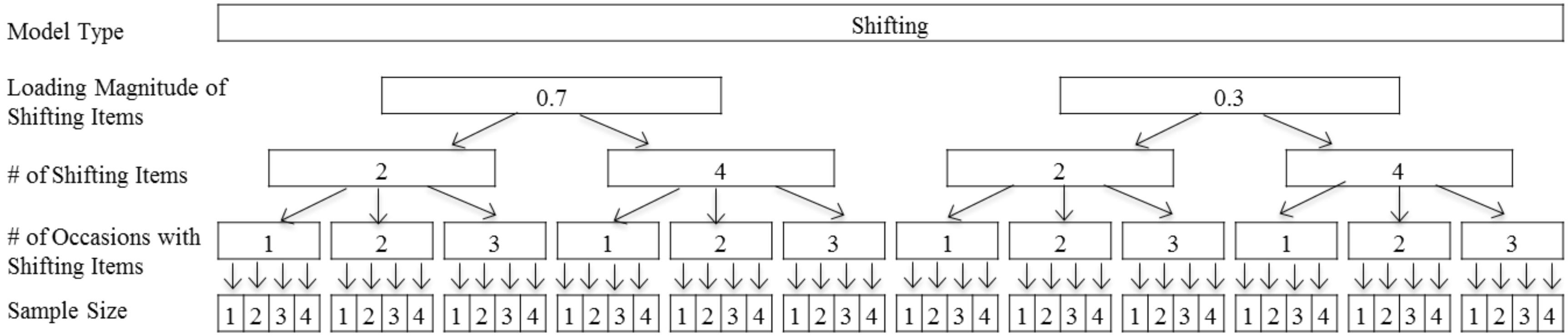


Figure 3.6. The 48 cells in the shifting model condition. All datasets were initially generated with all items present at each time point. Under the shifting indicators model, items were dropped according to the number of items (two items or four items) and number of occasions (one, two, or three occasions) conditions, respectively. Sample size: 1 = 250, 2 = 500, 3 = 750, 4 = 1000.

One other aspect of the study design warrants further description. Longitudinal equality constraints were imposed on the factor loadings of most items. In the case of the shifting indicators model, all item loadings for corresponding items were constrained to be equal over time (e.g., item 1 at time 1, item 1 at time 2, item 1 at time 3). In the case where items are dropped (e.g., if item 1 at time 1 is dropped), the remaining items were still constrained to be equal (e.g., item 1 at time 2, item 1 at time 3). Within the full model condition, equality constraints were also placed on the loadings of corresponding items in order to match the constraints on the shifting indicators model. In the case where the magnitude of the factor loadings of the shifting items condition took the value of .7, all item loadings were constrained to be equal to corresponding items at each time point. That is, full longitudinal measurement invariance at the metric level was imposed. In the case where the magnitude of the factor loadings of the shifting items condition took the value of .3, all items with factor loadings of .7 were constrained to be equal, but items with a factor loading of .3 were released. This is because constraining corresponding items with radically different factor loadings (i.e., .3 and .7) to be equal over time would represent an improper constraint. Therefore, partial longitudinal measurement invariance at the metric level was imposed. Note that there was only one cell where .3 loadings could be constrained to be equal over time because under all other conditions, corresponding items did not have .3 loadings at adjoining time points. In the one case where four items shifted across three time points, there were four items that had .3 loadings at adjoining time points. However, in this case, the .3 loadings were not constrained to be equal due to theoretical reasons. Specifically, it was assumed that in applied settings, items that had low loadings across more than one time point would not necessarily have the *same* low loading over time.

CHAPTER 4

RESULTS

This chapter presents the results of the Monte Carlo simulation comparing the performance of two types of second-order LGMs: (1) the second-order LGM with shifting indicators, and (2) a second-order LGM with all items present at each time point. Results are reported for each model type on the dependent variables in the following order: (1) model convergence to a proper solution, (2) growth parameter bias, (3) growth parameter standard error bias, (4) efficiency, and (5) model fit.

Convergence to a Proper Solution

Model convergence to a proper solution was investigated with regard to two outcomes: (1) failure to converge, and (2) convergence to an improper solution (i.e., involving a non-positive definite variance/covariance matrix).

With regard to model convergence, the full model and the shifting indicators model both performed well. Under the full model conditions, all of the 28,000 simulated datasets converged, for a convergence rate of 100%. Under the shifting model conditions, 47,986 of the 48,000 simulated datasets converged, yielding a convergence rate of 99.97%. The overall convergence rate across both conditions was 99.98%.

Table 4.1 presents the distribution of the 14 datasets that failed to converge. Of the 14 datasets that failed to converge, eight occurred when *n* = 250; however, within that condition, the convergence rate was still extremely high (11,992/12,000 = 99.93%). Half of the datasets that failed to converge had items dropped at each of the three measurement occasions.

With regard to convergence to solutions with a non-positive definite variance covariance matrix, the performance of the full model and the shifting indicators model was similar. Overall, both models performed well. Across the full model conditions, 27,930 of the 28,000 replications (99.75%) converged to a proper solution, and across the shifting model conditions, 47,822 of the 47,986 converged replications (99.65%) produced a proper solution. In total, 234 replications produced improper solutions, yielding an overall proper convergence rate of 99.67%. In all cases, the out-of-range value that caused the non-positive definite variance/covariance matrix was a negative slope variance.

An inverse relationship existed between sample size and the number of improper solutions as presented in Table 4.1. The majority of improper solutions (94%) were obtained when *n* = 250. An additional 6% were obtained when *n* = 500. When *n* = 250, the number of improper solutions increased as more shifting items were included in the model across more time points.

The shifting indicators models tended to have slightly more improper solutions than the corresponding full models. The largest number of improper solutions was obtained when the sample size was 250, the factor loading magnitude of the shifting indicators was .3, the number of shifting items was 4, and the number of time points with shifting indicators was 3. Under these conditions, 17 improper solutions were produced under the full model, and 25 were produced under the shifting indicators model. This result indicates that dropping items with low loadings rather than retaining them may decrease the likelihood that the model will properly converge at small sample sizes, although the magnitude of the difference in convergence rates for these cells was small.

Table 4.1

Number of Improper (and Nonconverged) Solutions Obtained Across All Conditions

		Full Model			Shifting Model			
		.3		.7	.3		.7	
		Number of Shifting Items		Number of Shifting Items	Number of Shifting Items		Number of Shifting Items	
Number of Times with Shifting Items	*N*	2	4	0	2	4	2	4
0	250	--	--	10	--	--	--	--
	500	--	--	0	--	--	--	--
	750	--	--	0	--	--	--	--
	1000	--	--	0	--	--	--	--
1	250	6	6	--	8	6	11	10 (1)
	500	1	0	--	1	0	1	1
	750	1	0	--	0	0	0	0 (1)
	1000	1	0	--	0	0	0 (3)	0
2	250	9	10	--	12	17 (1)	12	14
	500	0	0	--	0	0	1	3
	750	0	0	--	0	0	0	0 (1)
	1000	0	0	--	0	0	0	0
3	250	10	17	--	8	25 (3)	12 (3)	17
	500	0	1	--	0	2 (1)	1	2
	750	0	0	--	0	0	0	0
	1000	0	0	--	0	0	0	0

Note. The number of nonconverged solutions within each cell (if any) appears in parentheses. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the number of improper/nonconverged solutions under the loading magnitude of .7 is displayed only once per sample size condition for the full model. The values for shifting models under the loading magnitude of .7 can all be compared to those four values.

Considering all inadmissible solutions (i.e., nonconverged and improper solutions together), the performance of both the full and the shifting indicators model was good. Across all cells, the overall rate of convergence to a proper solution exceeded the criterion value of 90%. The lowest convergence rate (97.2%) was obtained under the shifting indicators model when four items of loading magnitude .3 were dropped in all three measurement occasions. Under these conditions, if the corresponding full model had been selected by a (hypothetical) researcher, the convergence rate would have been slightly higher (i.e., 98.3%), but the magnitude of this difference is small.

One consideration in Monte Carlo studies is whether to retain or replace replications that did not converge or failed to properly converge. Although parameter estimates obtained via improper solutions may not be trustworthy, some researchers may still use them despite receiving a warning message. Therefore, it can be informative to report results of the simulation with all solutions retained as well as with inadmissible solutions removed and replaced. In this study, results are presented both ways in the section summarizing bias in the parameter estimates. In this case, the results were somewhat different between the two methods of analysis. For all other results, no substantive differences were observed between the two methods, so results are only reported for the analyses that retained the inadmissible solutions.

Bias in the Parameter Estimates

Relative bias in the growth parameter estimates (i.e., slope mean, intercept mean, slope variance, intercept variance, slope/intercept covariance) was calculated and compared across the full model and shifting indicators model.

Values of relative bias of the growth parameter estimates for both the full models and the shifting indicator models were acceptable under all conditions (i.e., did not exceed .05, Hoogland & Boomsma, 1998) when the inadmissible solutions were retained. The slope mean and intercept mean estimates had less bias than slope variance, intercept variance, and slope/intercept covariance, but in all cases the amount of bias was small. Bias in the parameters was largest when the sample size was 250, as displayed in Table 4.2. It was expected that bias would systematically decrease as sample size increased. However, this result was only observed for the slope mean. The fluctuations observed in the other bias values (e.g., a larger slope variance bias at n = 750 than n = 500) appear to represent random sampling error around zero, as the values are all very small and quite similar to each other.

Table 4.2
Mean Relative Bias of Parameter Estimates by Sample Size, Collapsed Across All Other Conditions, With Inadmissible Solutions Retained

N	Slope mean	Intercept mean	Slope variance	Intercept variance	Slope/intercept covariance
250	.0037	.0040	.0201	.0141	-.0190
500	.0022	.0006	.0077	.0037	.0003
750	.0017	.0012	.0085	.0077	-.0124
1000	.0010	-.0007	.0037	.0042	-.0040

Relative bias of the parameter estimates was similar across the full and shifting indicators models. Table 4.3 presents mean values of relative bias collapsed across sample size. The full model and shifting model values are presented side by side for ease of comparison. When the loading magnitude of the shifting items was low (i.e., .3), there was very little difference between the parameter bias in the shifting models (which dropped those items) and the full models (which retained those items). In the case where the loading magnitude of the shifting items was high (i.e., .7), the parameter bias in the full model was comparable to the shifting models under all circumstances for the slope mean and intercept mean. For the slope variance, intercept variance, and slope/intercept covariance, the shifting models had larger values of bias (in an absolute value sense) as more items were dropped from more time points. In all cases, however, the amount of bias present in the estimates was within the acceptable limit.

Table 4.3

Mean Relative Bias of Parameter Estimates, Collapsed Across Sample Size, With Inadmissible Solutions Retained

			Slope mean		Intercept mean		Slope variance		Intercept variance		Slope/intercept covariance	
Loadings of shifting items	# times with shifting items	# shifting items	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting
.3	1	2	0.002	0.002	0.001	0.001	0.004	0.004	0.007	0.007	-0.003	-0.003
		4	0.003	0.003	0.000	0.000	0.005	0.005	0.006	0.006	-0.007	-0.007
	2	2	0.003	0.003	0.001	0.001	0.018	0.018	0.006	0.006	-0.009	-0.009
		4	0.002	0.002	0.002	0.002	0.013	0.012	0.005	0.006	-0.001	0.000
	3	2	0.003	0.003	0.001	0.001	0.007	0.008	0.007	0.007	-0.013	-0.013
		4	0.002	0.003	0.005	0.005	0.012	0.012	0.009	0.010	0.001	0.002
.7	0	0	0.001	--	0.001	--	0.006	--	0.007	--	-0.012	--
	1	2	--[a]	0.001	--	0.001	--	0.008	--	0.007	--	-0.013
		4	--	0.001	--	0.001	--	0.009	--	0.008	--	-0.013
	2	2	--	0.002	--	0.001	--	0.009	--	0.008	--	-0.014
		4	--	0.002	--	0.000	--	0.011	--	0.009	--	-0.016
	3	2	--	0.002	--	0.001	--	0.010	--	0.008	--	-0.014
		4	--	0.002	--	0.001	--	0.017	--	0.012	--	-0.022

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the full model bias value is displayed only once. The values for shifting models can all be compared to the same value.

[a] Elements of the table marked with a dash (--) indicate combinations of conditions for which the full model is the same.

The same set of analyses was run with extra replications to replace those that did not converge to a proper solution. Table 4.4 displays the relative bias of the parameter estimates reported by sample size.

Table 4.4

Mean Relative Bias of Parameter Estimates by Sample Size, Collapsed Across All Other Conditions, With Inadmissible Solutions Replaced

N	Slope mean	Intercept mean	Slope variance	Intercept variance	Slope/intercept covariance
250	.0046	.0049	.0334	.0176	-.0269
500	.0022	.0007	.0085	.0040	-.0003
750	.0017	.0012	.0085	.0077	-.0124
1000	.0010	-.0007	.0037	.0042	-.0040

Similar to the results summarized above, the relative bias in the parameter estimates was largest (in an absolute value sense) when the sample size was 250. For the slope mean, the bias values systematically decreased as sample size increased. For the other parameter estimates, bias did not systematically decrease with sample size. As in the Table 4.2, the magnitude of these fluctuations was small and likely represents random sampling error.

Table 4.5 displays the relative bias of the parameter estimates collapsed across sample size. The same general trends were observed when the inadmissible solutions were replaced as when they were retained. In comparing the full models to the corresponding shifting models, the amount of bias was similar between the shifting models that dropped items with low loadings and the full models that retained those items. In the case where the shifting items had high loadings, the bias in the slope mean and intercept mean was comparable across the full and shifting models. The bias in the slope variance, intercept variance, and slope/intercept covariance tended to be larger (in an absolute value sense) for the shifting models than the full models, especially as more items were dropped from more measurement occasions.

Table 4.5

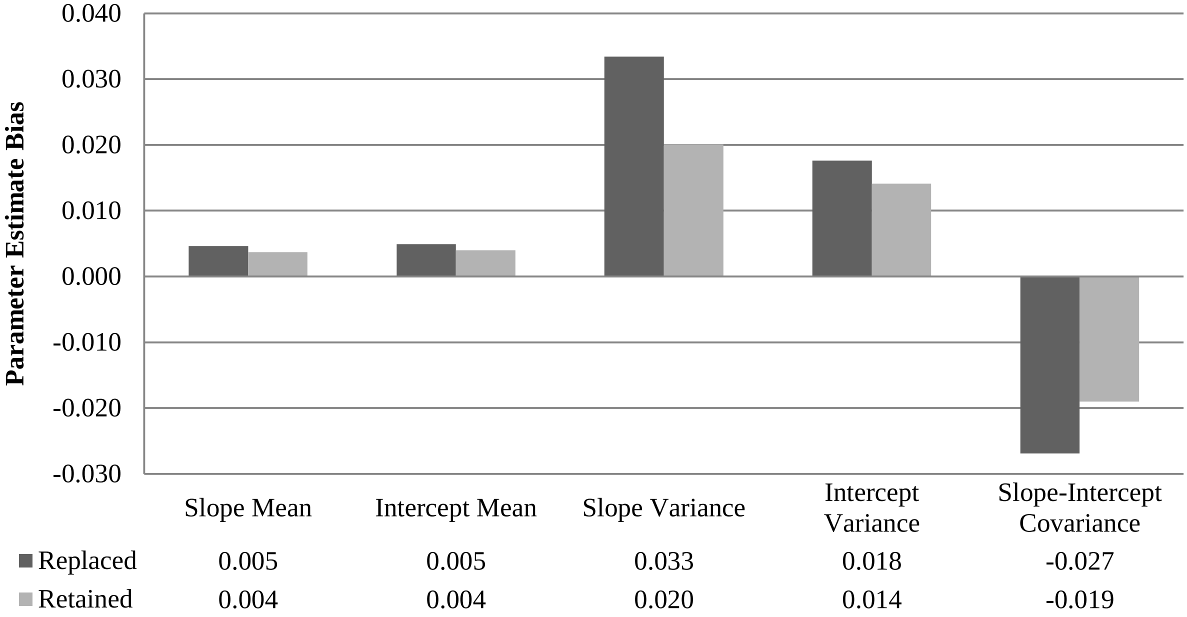
Mean Relative Bias of Parameter Estimates, Collapsed Across Sample Size, With Inadmissible Solutions Replaced

Loadings of shifting items	# times with shifting items	# shifting items	Slope mean		Intercept mean		Slope variance		Intercept variance		Slope/intercept covariance	
			Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting
.3	1	2	0.002	0.002	0.001	0.001	0.006	0.006	0.008	0.008	-0.004	-0.004
		4	0.003	0.003	0.000	0.000	0.006	0.007	0.006	0.006	-0.008	-0.009
	2	2	0.004	0.004	0.001	0.001	0.020	0.021	0.007	0.006	-0.011	-0.010
		4	0.002	0.003	0.003	0.003	0.017	0.018	0.007	0.008	-0.004	-0.004
	3	2	0.003	0.003	0.001	0.001	0.011	0.010	0.008	0.008	-0.015	-0.015
		4	0.003	0.003	0.005	0.006	0.017	0.020	0.010	0.012	-0.002	-0.003
.7	0	0	0.001	--	0.001	--	0.009	--	0.008	--	-0.013	--
	1	2	-- [a]	0.002	--	0.001	--	0.011	--	0.008	--	-0.015
		4	--	0.002	--	0.001	--	0.012	--	0.008	--	-0.015
	2	2	--	0.002	--	0.001	--	0.013	--	0.009	--	-0.016
		4	--	0.002	--	0.001	--	0.016	--	0.010	--	-0.018
	3	2	--	0.002	--	0.001	--	0.014	--	0.009	--	-0.015
		4	--	0.002	--	0.002	--	0.023	--	0.013	--	-0.024

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the full model bias value is displayed only once. The values for shifting models can all be compared to the same value.

[a] Elements of the table marked with a dash (--) indicate combinations of conditions for which the full model is the same.

The biggest differences in the bias values once inadmissible solutions were replaced occurred when the sample size was 250, as displayed in Figure 4.1. It is expected that these values would differ the most because the majority of inadmissible solutions (96%) occurred when *n* = 250. In terms of the parameter estimates, the most substantial difference was observed in the slope variance. This result is expected because negative slope variance was the cause of all of the improper solutions; therefore, these values changed the most as out-of-range estimates were replaced with proper solutions. The slope variance bias estimate was slightly smaller when inadmissible solutions were retained than when they were replaced. The reasons behind this change are discussed in more detail below.



	Slope Mean	Intercept Mean	Slope Variance	Intercept Variance	Slope-Intercept Covariance
■ Replaced	0.005	0.005	0.033	0.018	-0.027
■ Retained	0.004	0.004	0.020	0.014	-0.019

Figure 4.1. Relative bias of the growth parameter estimates at sample size = 250 with inadmissible solutions retained and replaced.

Although, the parameter estimate bias results were generally similar with or without the inadmissible solutions, in two cases, the results were made substantively worse by replacing inadmissible solutions. In the shifting model with four dropped items at three time points and sample size of 250, the mean relative bias for the slope variance exceeded the criteria for acceptable bias of .05. This result occurred under both loading magnitude conditions. When the loading magnitude of the shifting items was .3, the mean relative bias for this cell was .054. When the loading magnitude of the shifting items was .7, the mean relative bias was .052. When the inadmissible solutions were retained in the data, the mean relative bias for these cells was .024 and .031, respectively.

To explain this change, a closer examination of the parameter estimates for one of the cells is instructive. Figure 4.2 presents the histogram of slope variance estimates for the cell with 3 items of loading magnitude of .3 dropped at three time points. In this figure, observations to the left of the vertical line represent slope variance estimates that are negative and inadmissible. (Note: The true value of the slope variance is .2.) When these observations are discarded and replaced by admissible solutions, as in Figure 4.3, the left tail of the distribution is eliminated. The average slope variance bias for the improper solutions in this cell was -.028. By removing these values and replacing them with admissible parameter estimates, the mean of the parameter estimate within the cell shifted from .205 to .211. Therefore, the distribution of the admissible solutions eliminated out-of-range values, the mean value was pulled to the right, and the cell mean was less accurate.

All improper solutions were caused by negative slope variance, so this parameter is affected more than others by the removal of inadmissible solutions. This shift in the means of the slope variance occurred across all cells (unless there were no inadmissible solutions to replace), but the effect was more pronounced for these two particular cells, which had the largest number of inadmissible solutions in the study. Aside from these cases, the relative bias of the parameter estimates for all cells was less than .05.

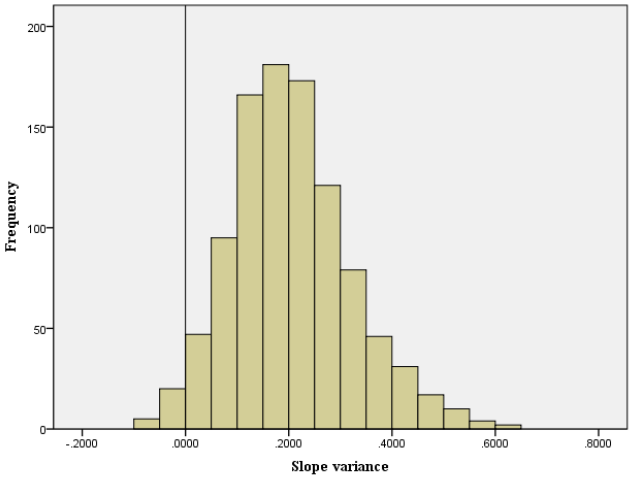


Figure 4.2. Slope variance estimates for the shifting model with four dropped items of loading magnitude .3 at three time points and a sample size of 250. Inadmissible solutions are retained. Observations to the left of the vertical line represent negative slope variance estimates and are improper.

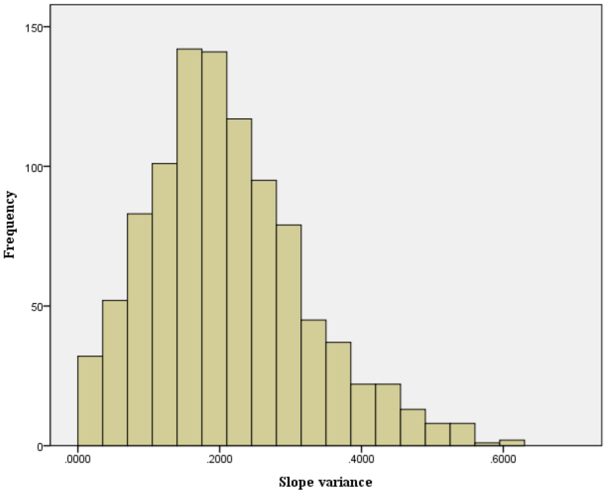


Figure 4.3. Slope variance estimates for the shifting model with four dropped items of loading magnitude .3 at three time points and a sample size of 250. Inadmissible solutions are removed, and 28 replacement observations were generated to replace three runs that failed to converge and 25 runs with improper solutions.

ANOVAs were conducted to investigate whether there were any significant differences in relative bias between the levels of the independent variables with a partial eta-squared value of greater than .05. For each dependent variable, ANOVAs were run with inadmissible solutions retained and with inadmissible solutions dropped and replaced with additional properly converged replications. Five-way ANOVAs (model type x loading magnitude x number of items x number of times x sample size) were conducted including all independent variables as factors. However, there are some drawbacks to this method of testing. Because the study design is not completely crossed (see Figures 3.4 and 3.5), there are large differences in sample size across levels of some independent variables. For example, there are 28,000 cases in the full model condition and 48,000 cases in the shifting model condition. Unequal sample sizes can be problematic in ANOVA when the equal variance assumption is not met (Keppel & Wickens, 2004, p. 149). For all ANOVAs conducted in this study, the equal variance assumption was not met (i.e., Levene's test was significant); a main reason for this is that the variance in the smaller sample size conditions are systematically larger than the variance in the larger sample size conditions. The combination of unequal variances and unequal sample sizes renders any conclusions drawn from these ANOVAs tentative.

Results of the ANOVAs indicated that for all growth parameter estimates there were no differences that exceeded the .05 partial eta-squared criterion. These results seem reasonable in light of the small differences in means presented in Tables 4.2 – 4.5.

Bias in the Standard Errors of the Parameter Estimates

Relative bias in the standard errors of the growth parameter estimates was calculated and compared across the full and shifting indicators models. There were no substantive differences between the sets of results with inadmissible solutions retained or replaced. Results are reported only for the set of results with inadmissible solutions retained.

The relative bias of the standard errors of the growth parameter estimates was acceptable under all conditions (i.e., did not exceed .1). Table 4.6 presents mean values of relative bias reported by sample size. For the slope variance and intercept variance, the relative bias in the standard errors was largest when *n* = 250. Otherwise, there did not seem to be a discernible trend relating sample size to mean relative bias in the standard errors. Given that all of the values are quite small, it is likely that the small fluctuations seen in the relative bias of the standard errors represent random sampling error around zero.

Table 4.6
Mean Relative Bias of the Standard Errors Of Growth Parameter Estimates by Sample Size

N	Slope mean	Intercept mean	Slope variance	Intercept variance	Slope/intercept covariance
250	-.0091	.0038	-.0144	-.0164	.0048
500	.0174	-.0076	-.0003	-.0030	.0004
750	-.0114	-.0159	-.0078	.0136	.0054
1000	.0108	-.0048	-.0104	-.0119	-.0080

Table 4.7 presents mean values of relative bias of the standard errors collapsed across sample size. For the models where the loading magnitude of the shifting items was .3, results were generally similar between the full models that retained those items and the shifting models that dropped those items. For the slope variance, intercept variance, and slope/intercept covariance, the magnitude of the difference of the bias values between the shifting and full models grew when four items were dropped at each of the three measurement occasions. Even in those cases, however, the overall amount of bias was within the acceptable limit.

Table 4.7

Mean Relative Bias of the Standard Errors of Growth Parameter Estimates, Collapsed Across Sample Size

			Slope mean		Intercept mean		Slope variance		Intercept variance		Slope/intercept covariance	
Loadings of shifting items	# times with shifting items	# shifting items	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting
.3	1	2	-0.015	-0.016	0.004	0.004	-0.018	-0.018	-0.025	-0.025	-0.014	-0.015
		4	-0.003	-0.003	0.003	0.005	-0.003	-0.004	0.011	0.012	0.004	0.004
	2	2	0.004	0.004	-0.003	-0.001	0.000	0.001	0.014	0.013	-0.003	-0.002
		4	-0.006	-0.007	-0.005	-0.003	0.010	0.012	0.013	0.014	0.008	0.011
	3	2	0.010	0.009	0.003	0.004	-0.003	-0.005	0.000	0.000	-0.006	-0.005
		4	0.006	0.001	0.002	-0.004	0.001	-0.010	-0.009	-0.016	-0.005	-0.010
.7	0	0	0.010	--	-0.020	--	-0.014	--	-0.012	--	0.006	--
	1	2	--[a]	0.009	--	-0.020	--	-0.013	--	-0.013	--	0.008
		4	--	0.007	--	-0.019	--	-0.015	--	-0.012	--	0.009
	2	2	--	0.009	--	-0.021	--	-0.014	--	-0.013	--	0.006
		4	--	0.009	--	-0.017	--	-0.022	--	-0.014	--	0.004
	3	2	--	0.008	--	-0.018	--	-0.015	--	-0.012	--	0.006
		4	--	0.000	--	-0.012	--	-0.026	--	-0.010	--	0.004

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the full model bias value is displayed only once. The values for shifting models can all be compared to the same value.

[a] Elements of the table marked with a dash (--) indicate combinations of conditions for which the full model is the same.

For the models where the loading magnitude of the shifting items was .7, results were generally comparable between the full model that retained those items and the shifting models that dropped those items. Within the shifting models, a clear pattern did not emerge relating changes in the bias values to the number of dropped items per factor or the number of measurement occasions when items were dropped.

The five-way ANOVA results indicated that there were no differences in relative bias of the standard errors that exceeded to partial eta-squared criterion of .05. As described earlier, these results should be interpreted tentatively due to the violation of the assumption of equal variances. However, the results seem reasonable in light of the small differences in means presented in Tables 4.6 – 4.7.

Efficiency of the Parameter Estimates

Efficiency of the growth parameter estimates was compared across the full and shifting indicators models. Smaller values are desirable because efficiency was measured as the average standard error of the growth parameter estimates. Table 4.8 presents average standard errors of the growth parameters by sample size. As would be expected, average standard errors decreased as sample size increased. In all cases, the magnitude of the difference in efficiency between *n* = 250 and *n*= 500 was greater than the magnitude of the difference between the other levels of sample size.

Table 4.8
Average Standard Errors of the Growth Parameter Estimates by Sample Size

N	Slope mean	Intercept mean	Slope variance	Intercept variance	Slope/intercept covariance
250	0.079	0.104	0.103	0.265	0.092
500	0.056	0.074	0.073	0.187	0.065
750	0.046	0.060	0.060	0.153	0.053
1000	0.039	0.052	0.051	0.132	0.046

Table 4.9 presents the average values of the standard errors of the growth parameters collapsed across sample size. For the models where the loading magnitude of the shifting items was .3, results were generally comparable between the full models that retained those items and the shifting models that dropped those items.

For the models where the loading magnitude of the shifting items was .7, the full model generally had lower average standard errors (and thus greater efficiency) than the shifting models. These values were similar when two items were dropped from one time point, but grew more disparate as more items were dropped from more time points.

Table 4.9

Average Standard Errors of Growth Parameter Estimates, Collapsed Across Sample Size

			Slope mean		Intercept mean		Slope variance		Intercept variance		Slope/intercept covariance	
Loadings of shifting items	# times with shifting items	# shifting items	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting
.3	1	2	0.053	0.053	0.071	0.071	0.069	0.069	0.178	0.178	0.061	0.061
		4	0.054	0.054	0.071	0.071	0.070	0.071	0.179	0.179	0.062	0.062
	2	2	0.053	0.053	0.071	0.071	0.070	0.070	0.179	0.180	0.062	0.062
		4	0.056	0.056	0.073	0.073	0.072	0.073	0.185	0.186	0.065	0.066
	3	2	0.054	0.054	0.072	0.072	0.071	0.071	0.182	0.183	0.062	0.063
		4	0.061	0.062	0.077	0.079	0.077	0.080	0.203	0.209	0.072	0.075
.7	0	0	0.052	--	0.071	--	0.069	--	0.177	--	0.061	--
	1	2	--[a]	0.053	--	0.071	--	0.069	--	0.178	--	0.061
		4	--	0.054	--	0.071	--	0.071	--	0.179	--	0.062
	2	2	--	0.053	--	0.072	--	0.070	--	0.180	--	0.062
		4	--	0.056	--	0.073	--	0.073	--	0.186	--	0.066
	3	2	--	0.054	--	0.072	--	0.071	--	0.183	--	0.063
		4	--	0.060	--	0.076	--	0.076	--	0.199	--	0.069

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the full model standard error value is displayed only once. The values for shifting models can all be compared to the same value.

[a] Elements of the table marked with a dash (--) indicate combinations of conditions for which the full model is the same.

Results of the ANOVAs indicated that the effect of several of the independent variables on efficiency exceeded the partial eta-squared criterion of .05. These values are presented in Table 4.10.

Table 4.10
Partial Eta-Squared Values Related to Efficiency that Exceeded .05

Growth Parameter	Sample Size	# Times	# Items	Times x Items
Slope mean	.94	.20	.19	.12
Intercept mean	.94	.09	.06	--
Slope variance	.77	--	--	--
Intercept variance	.77	.05	--	--
Slope-intercept covariance	.78	.06	.06	--

The most substantial effect was of sample size on efficiency. Figure 4.4 displays the change in efficiency values for each of the growth parameters as sample size increases. As would be expected, efficiency improves (i.e., the average standard errors decrease) as sample size increases.

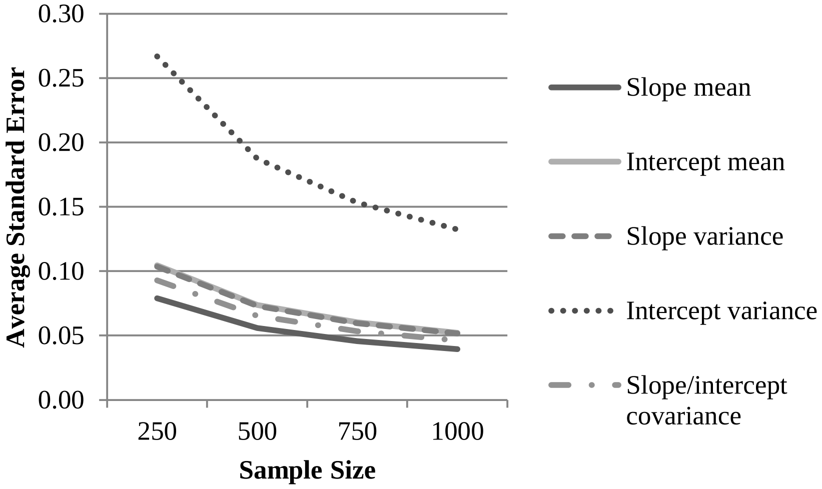


Figure 4.4. Average standard errors of the growth parameters by sample size.

The number of measurement occasions when items were shifted had a smaller effect on the efficiency of four growth parameters: slope mean, intercept mean, intercept variance, and slope/intercept covariance. Figure 4.5 displays the change in efficiency in these parameters as the number of times increases. For each of the growth parameters, a small increase in the average standard errors of the parameter estimates accompanied an increase in the number of times when items were shifted, indicating a slight decline in efficiency. For the slope variance, the overall trend was the same (i.e., larger standard errors were observed as the number of times with shifting indicators increased), but the effect did not exceed the partial eta-squared criteria of .05.

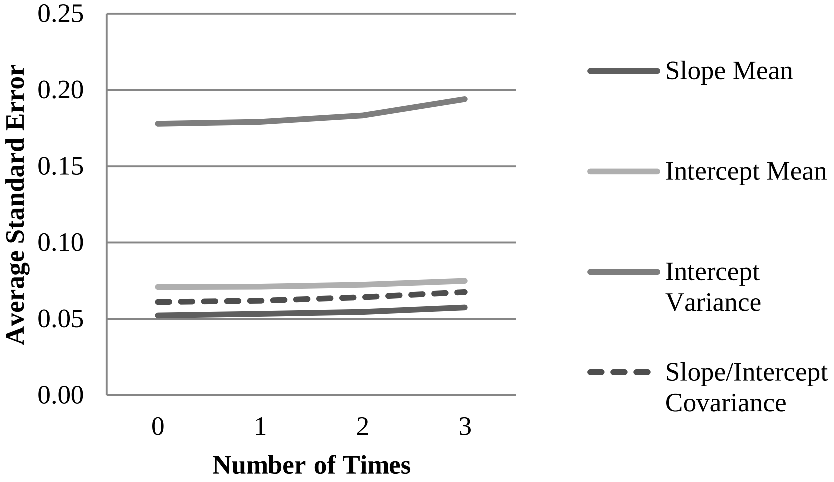


Figure 4.5. Average standard errors of the growth parameters by number of times which included shifted items.

The number of items variable also had a small effect on the efficiency of three of the growth parameters: slope mean, intercept mean, and slope/intercept covariance. Figure 4.6 displays the change in efficiency as the number of items that are shifted increased. As the number of shifted items increased, the average standard errors of the parameter estimates became larger, indicating a reduction of efficiency. For the slope and intercept variance, average standard errors also increased as the number of shifting items increased, but the effect did not exceed the partial eta-squared criteria of .05.

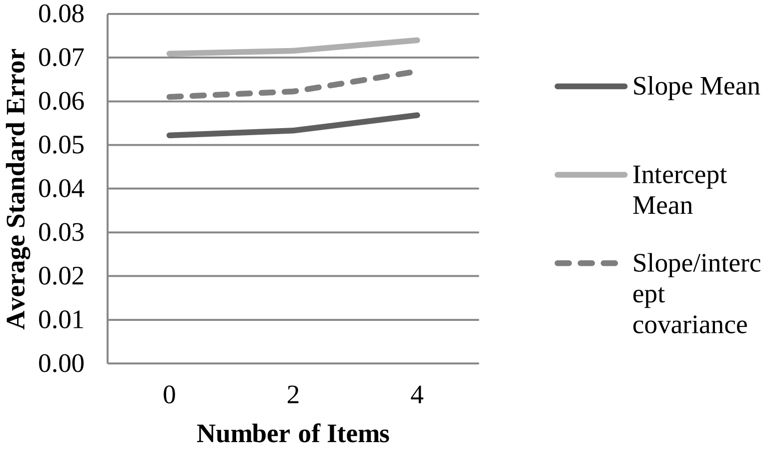


Figure 4.6. Average standard errors for the growth parameters by number of shifted items.

For the slope mean, the interaction of the number of times with shifted indicators and the number of shifted items per occasion variables exceeded the partial eta-squared criterion of .05. Figure 4.7 displays this interaction effect. When the number of shifting items was 4, there was a much larger increase in average standard errors as the number of occasions with shifted items increased than when the number of shifting items was 2.

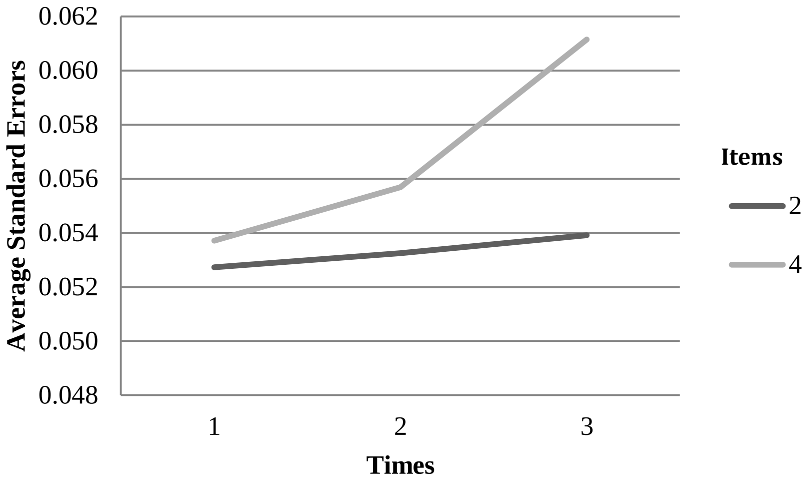


Figure 4.7. The interaction of the number of times and the number of items on the average standard errors of the slope mean.

Model Fit

Several measures of model fit were compared across the full and shifting indicators models. For every cell in the study design, the proportion of Type I errors in the chi-square fit statistic, and the average values of CFI, TLI, and RMSEA (and the upper bound of its confidence interval) were calculated.

The chi-square rejection rate was investigated across each cell. Within each cell, the percentage of replications that would have been rejected (p < .05) by the chi-square test (i.e., Type I errors) were calculated by comparing the observed chi-square fit statistic to the expected chi-square value based on the degrees of freedom of the tested model. Table 4.11 presents the percentage of replications with sample chi-square values that exceeded the critical value within the cell. For the cells with a sample size of 250, the Type I error rate was highest. Of the 19 cells in the *n*=250 condition, 14 cells (73.7%) had a Type I error rate of greater than 7.5%, the upper bound of Bradley's (1978) liberal criterion. The full model condition tended to have higher Type I error rates than the shifting indicator models. Within the full model condition, 14 out of 28 cells (i.e., 50%) exceeded the criterion, in comparison to 7 of 48 cells (i.e., 14.6%) in the shifting indicators condition. Within the shifting indicators condition, the Type I error rate tended to improve as the number of omitted items increased. This result indicates that the improvement in the rejection rate is likely being driven by the reduction of indicators per factor as items are dropped. The covariance matrix associated with the shifting indicators model is smaller than the covariance matrix for the full model, leading to fewer opportunities for discrepancies between the original and reproduced covariance matrices. This effect (i.e., higher rates of rejection for models with more indicators per factor) was previously reported for CFA models by Marsh, Hau, Balla, and Grayson (1998) and second-order LGMs by Leite (2007).

Table 4.11

Percentage of Type I Errors According to the Chi-Square Statistic Across All Conditions

		Full Model			Shifting Model			
		.3		.7	.3		.7	
		Number of Items		Number of Items	Number of Items		Number of Items	
Number of Times	*N*	2	4	0	2	4	2	4
0	250	--	--	9.7	--	--	--	--
	500	--	--	8.6	--	--	--	--
	750	--	--	7.3	--	--	--	--
	1000	--	--	6.5	--	--	--	--
1	250	11.8	10.6	--	9.4	9.3	9.5	8.3
	500	7.2	7.8	--	6.5	6.6	7.3	7.4
	750	7.0	6.9	--	6.0	6.5	6.8	5.5
	1000	6.2	6.5	--	5.2	6.7	6.1	6.3
2	250	12.9	13.4	--	11.1	7.5	9.2	6.8
	500	8.3	6.8	--	6.5	5.9	7.4	6.0
	750	8.8	8.0	--	7.4	6.0	7.2	5.4
	1000	7.0	5.2	--	6.5	4.3	5.5	5.0
3	250	11.9	12.9	--	7.0	7.4	8.7	7.1
	500	7.0	8.2	--	5.9	5.5	7.5	6.5
	750	6.4	6.4	--	5.7	5.8	5.8	4.9
	1000	6.7	6.8	--	6.9	6.1	6.2	5.2

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the percentage of Type I errors under the loading magnitude of .7 is displayed only once per sample size condition under the full model. The values for shifting models under the loading magnitude of .7 can all be compared to those four values.

With regard to CFI and TLI, the average values obtained by the full and shifting models indicated that the fit of the model was good. Values of CFI and TLI increased as sample size increased, as displayed in Table 4.12, but even with the smallest sample size the fit indices exceeded the recommended cutoff of .95.

Table 4.12

Mean Values of Fit Indices by Sample Size

N	CFI	TLI	RMSEA	RMSEA Upper
250	.997	.999	.011	.029
500	.999	.999	.006	.020
750	.999	1.00	.005	.016
1000	.999	1.00	.005	.014

RMSEA Upper: The upper bound of the confidence interval associated with the value of RMSEA.

Table 4.13 presents the average values of the CFI and TLI collapsed across sample size. Very little variation in these values was observed across the levels of independent variables, with all values falling within the range of .998-1.00. For the full models, the CFI decreased slightly as more items with loadings of .3 were included in the model. For shifting models, the TLI increased slightly as more items were dropped from more time points.

Table 4.13

Average Values of Fit Indices, Collapsed Across Sample Size

			CFI		TLI		RMSEA		RMSEA Upper	
Loadings	Times	Items	Full	Shifting	Full	Shifting	Full	Shifting	Full	Shifting
.3	1	2	0.999	0.999	0.999	0.999	0.007	0.007	0.018	0.018
		4	0.998	0.999	0.999	1.000	0.007	0.007	0.017	0.019
	2	2	0.998	0.999	0.999	0.999	0.007	0.007	0.018	0.019
		4	0.998	0.999	0.999	1.000	0.007	0.007	0.018	0.021
	3	2	0.998	0.999	0.999	1.000	0.007	0.007	0.018	0.020
		4	0.997	0.999	0.999	1.000	0.007	0.008	0.018	0.026
.7	0	0	0.999	--	0.999	--	0.007	--	0.018	--
	1	2	--[a]	0.999	--	0.999	--	0.007	--	0.018
		4	--	0.999	--	0.999	--	0.007	--	0.019
	2	2	--	0.999	--	0.999	--	0.007	--	0.019
		4	--	0.999	--	1.000	--	0.007	--	0.021
	3	2	--	0.999	--	1.000	--	0.007	--	0.020
		4	--	0.999	--	1.000	--	0.008	--	0.025

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the full model fit index value is displayed only once. The values for shifting models can all be compared to the same value.

[a] Elements of the table marked with a dash (--) indicate combinations of conditions for which the full model is the same.

RMSEA Upper: The upper bound of the confidence interval associated with the value of RMSEA

Rates of “rejection” for the CFI and TLI were calculated for each cell. (The term “rate of rejection” is used in the interest of simplicity to indicate that the fit value did not meet the commonly-accepted criterion for acceptable fit of .95. Because the CFI and TLI are descriptive measures of fit with recommended cut-off criteria, rather than tests of statistical inference like the chi-square fit statistic, the rate of rejection is not analogous to a Type I error rate.) Table 4.14 displays the percentage of rejected models within each level of the independent variables for the CFI, and Table 4.15 displays this information for the TLI. The highest rejection rates were observed when the sample size was 250. For both the CFI and the TLI, the rate of rejection when *n* = 250 was substantially higher than the comparable rate of rejection according to the chi-square fit statistic. Hu and Bentler (1999) noted that the TLI and CFI tended to over-reject true models at small sample sizes (≤ 250). At larger sample sizes, the CFI and TLI rates of rejection were generally lower than the chi-square Type I error rate.

For the full models where the loading magnitude of the shifting items was .3, the rate of rejection increased as more items at more time points were incorporated into the model with low loadings. The rate of rejection was especially high when four items at each of three measurement occasions had a loading magnitude of .3.

Overall, the rejection rates for CFI and TLI were lower for the shifting models than the corresponding full models. For the shifting models, as more items were dropped across more time points, the rejection rate generally decreased. This trend mirrors what was seen with the chi-square rates of rejection presented in Table 4.11. Again, this result seems to be driven by the reduction in the number of indicators per factor when items are dropped.

Table 4.14

Percentage of Rejected Models According to the CFI Across All Conditions

		Full Model			Shifting Model			
		.3		.7	.3		.7	
Number of		Number of Items		Number of Items	Number of Items		Number of Items	
Times	*N*	2	4	0	2	4	2	4
0	250	--	--	23.4	--	--	--	--
	500	--	--	2.9	--	--	--	--
	750	--	--	0.0	--	--	--	--
	1000	--	--	0.0	--	--	--	--
1	250	25.5	31.2	--	22.7	21.1	22.4	22.6
	500	3.6	5.3	--	2.5	3.0	2.6	3.7
	750	0.1	0.7	--	0.1	0.2	0.0	0.1
	1000	0.0	0.0	--	0.0	0.0	0.0	0.0
2	250	30.5	36.4	--	21.8	19.4	21.2	22.1
	500	5.1	8.2	--	3.1	3.0	3.1	3.7
	750	0.8	2.4	--	0.1	0.5	0.1	0.1
	1000	0.1	0.1	--	0.0	0.0	0.0	0.0
3	250	35.9	43.7	--	19.5	23.2	20.3	20.3
	500	6.0	19.5	--	2.6	5.5	2.4	4.4
	750	0.9	6.1	--	0.2	1.3	0.2	0.2
	1000	0.2	2.2	--	0.0	0.0	0.0	0.0

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the percentage of rejected models under the loading magnitude of .7 is displayed only once per sample size condition for the full model. The values for shifting models under the loading magnitude of .7 can all be compared to those four values.

Table 4.15

Percentage of Rejected Models According to the TLI Across All Conditions

		Full Model			Shifting Model			
		.3		.7	.3		.7	
Number of		Number of Items		Number of Items	Number of Items		Number of Items	
Times	*N*	2	4	0	2	4	2	4
0	250	--	--	21.9	--	--	--	--
	500	--	--	2.7	--	--	--	--
	750	--	--	0.0	--	--	--	--
	1000	--	--	0.0	--	--	--	--
1	250	25.1	30.7	--	21.2	20.1	20.9	21.4
	500	2.5	4.7	--	2.2	2.4	2.0	3.3
	750	0.1	0.5	--	0.0	0.1	0.0	0.1
	1000	0.0	0.0	--	0.0	0.0	0.0	0.0
2	250	29.9	36.0	--	20.9	18.6	19.9	21.2
	500	4.9	8.0	--	2.3	2.7	2.7	3.5
	750	0.8	2.3	--	0.1	0.4	0.1	0.1
	1000	0.1	0.1	--	0.0	0.1	0.0	0.0
3	250	35.1	43.8	--	18.6	23.6	19.6	20.7
	500	5.8	19.5	--	2.3	5.8	2.1	4.7
	750	0.8	6.1	--	0.1	1.3	0.2	0.2
	1000	0.2	2.2	--	0.0	0.1	0.0	0.0

Note. For the conditions with .3 loadings of the shifting items, there is a different full model for each shifting condition (i.e., depending on the number of times with shifting items and the number of shifting items per time point). For the conditions with .7 loadings of the shifting items, there is only one full model, regardless of the levels of the other shifting conditions. Because of this, the percentage of rejected models under the loading magnitude of .7 is displayed only once per sample size condition for the full model. The values for shifting models under the loading magnitude of .7 can all be compared to those four values.

With regard to RMSEA, the average values and confidence intervals indicated that model fit was good under the full and shifting indicators models. Under all conditions, the average value of RMSEA and the upper bound of the confidence interval was less than .06, as displayed in Table 4.12. As sample size increased, the values of RMSEA improved, as displayed in Table 4.13. However, the magnitude of the differences across sample size was quite small.

The mean RMSEA values were virtually identical across all conditions except sample size. However, the upper bound of the confidence interval differentiated between cells. For the models where the loading magnitude of the shifting items was .3, RMSEA upper bound values were similar for the full models that included these items and the shifting models that dropped them. However, for the models where the loading magnitude of the shifting items was .7, the RMSEA values became progressively higher (i.e., indicated worse fit) in the shifting models as more items were dropped from more time points. In contrast to the chi-square fit statistic, the CFI, and the TLI, the RMSEA did not improve as the number of items per factor decreased.

Rates of rejection based on the RMSEA and the upper bound of the RMSEA confidence interval were calculated. For the RMSEA, the rejection rate was 0.0% for all cells. That is, none of the replications in the simulation would have been rejected based on the value of RMSEA. Using the upper limit of the confidence interval for the RMSEA as the basis of the rejection rate, only two cells had a rejection rate higher than 0.0%. Both cells were in the shifting indicators condition with four items missing at each of the three time points and a sample size of 250. When the factor loadings of the dropped items had a magnitude of .3, the rejection rate was 2.3%. When the factor loadings of the dropped items had a magnitude of .7, the rejection rate was 1.8%.

Review of Research Hypotheses and Support from Findings

The purpose of this study was to compare the performance of second-order LGMs with shifting indicators to second-order LGMs with all items present at each time point. Four research questions were investigated. These questions focused on the impact of five independent variables related to the structure of the second order LGMs on five dependent variables related to model performance. Specifically, the research questions were:

1. How does the number of shifting to non-shifting indicators within the affected measurement occasions influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?
2. How does the number of measurement occasions with shifting indicators influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?
3. How does the magnitude of the factor loadings for the omitted items influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?
4. How does sample size influence model convergence, bias in growth parameter estimates, bias in standard error estimates, efficiency, and model fit?

This section restates the research hypotheses related to each of these questions and summarizes the findings in light of these hypotheses. The hypotheses generally state that the full models will outperform the shifting models (except in the case of model fit). This is due to the fact that the full models are properly specified and the shifting models omit items. Although it was expected that the full models would perform better than the shifting models, a critical consideration was the degree of difference in performance. If the shifting models performed profoundly worse than the full models, future use of the model may justifiably be limited to circumstances when no other alternative is available (i.e., items have been inadvertently lost). However, given that the performance of the two types of model was generally comparable, the flexibility gained by using the shifting model may outweigh any trivial loss in performance.

Hypothesis 1:

With respect to the number of shifting indicators within each measurement occasion, it was hypothesized that models with more indicators per factor within each measurement occasion would have better rates of proper convergence, less bias in parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation. With regard to fit, models with more indicators per factor within each measurement occasion were hypothesized to have larger model rejection rates based on the chi-square and related fit statistics.

Support for Hypothesis 1:

The findings of this study were generally in support of hypothesis 1. Models that had more indicators per factor (i.e., full models versus shifting models; shifting models with fewer dropped items per occasion versus shifting models with more dropped items per occasion) tended to have better rates of proper convergence and more efficient estimation. The overall magnitude of these differences, however, was small. Bias in the parameter estimates and standard errors was mostly comparable across the levels of the number of shifting items that were dropped. For selected growth parameters, bias in the parameter estimates and standard errors was larger when more items were dropped. Models with more indicators per factor within each measurement occasion had larger model rejection rates based on the chi-square fit statistic, CFI, and TLI, especially when the sample size was small (i.e., 250). Overall, the results indicated that having fewer indicators per factor did not have a substantial impact on model performance under the conditions tested in this simulation.

Hypothesis 2:

With respect to the number of measurement occasions that include shifting indicators, it was hypothesized that models with fewer occasions involving shifting indicators would have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation. With regard to fit, models with fewer occasions of shifting indicators were hypothesized to have larger model rejection rates based on the chi-square and related fit statistics.

Support for Hypothesis 2:

The findings of this study were generally in support of hypothesis 2. Models with fewer occasions involving shifting indicators (i.e., full models versus shifting models; shifting models with fewer occasions with dropped items versus shifting models with more dropped items) tended to have better rates of proper convergence and more efficient estimation. The overall magnitude of these differences was small. Bias in the parameter estimates and standard errors was mostly comparable across the levels of measurement occasions when items were dropped. For selected growth parameters, bias in the parameter estimates and standard errors rose as items were dropped at more measurement occasions, especially when four (versus two) items were dropped. Models with fewer occasions involving shifting indicators did have larger model rejection rates based on the chi-square and related fit statistics, especially when the sample size was small (i.e., 250). Overall, the results indicated that dropping items at multiple time points did not have a substantial impact on model performance under the conditions tested in this simulation. However, when four items were dropped at all three time points, proper convergence, bias, and efficiency became noticeably worse.

Hypothesis 3:

With respect to the magnitude of the factor loadings for the omitted items, it was hypothesized that models that drop items with low loadings would have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation than models that drop items with high loadings. With regard to fit, models that drop items with low loadings were hypothesized to have smaller model rejection rates based on the chi-square and related fit statistics.

Support for Hypothesis 3:

The findings of this study were mixed with regard to hypothesis 3. With regard to inadmissible solutions, the shifting models that dropped items with high loadings did have more inadmissible solutions than the shifting models that dropped items with low loadings. In terms of the relative bias of the parameter estimates, results were generally comparable between the shifting models that dropped items with low loadings and the shifting models that dropped items with high loadings. In terms of the relative bias of the standard errors of the parameter estimates, there were some differences between the two types of model, but a clear pattern was difficult to discern. Efficiency, as measured by the average value of standard errors of parameter estimates, was quite similar between the shifting models that dropped items with low loadings and the shifting models that dropped items with high loadings. Under certain conditions (i.e., when there was a large number of dropped items across two or three measurement occasions), the shifting models that dropped items with low loadings had slightly larger standard errors, indicating worse efficiency. With regard to measures of model fit, rejection rates tended to be generally comparable between the two types of models. Overall, the results indicated that dropping items with low versus high loadings did not have a substantial impact on model performance under the conditions tested in this simulation.

Hypothesis 4:

With respect to sample size, it was hypothesized that models with larger sample sizes would have better rates of proper convergence, less bias in the parameter estimates, less bias in the standard errors of the parameter estimates, and more efficient estimation than models with smaller sample sizes. With regard to fit, models with larger sample sizes were hypothesized to have smaller model rejection rates based on the chi-square and related fit statistics than models with smaller sample sizes.

Support for Hypothesis 4:

The findings of this study were generally supportive of hypothesis 4. Models with larger sample sizes had better rates of proper convergence and more efficient estimation. Bias in the parameter estimates was highest when *n* = 250. The relationship of sample size to bias in the standard errors was more complex. For the slope variance and intercept variance, the relative bias in the standard errors was largest when *n* = 250. Otherwise, no discernible pattern emerged. Models with larger sample sizes had smaller model rejection rates based on the chi-square and related fit statistics. The magnitude of these differences was most profound between the *n* = 250 condition and the *n* = 500 condition. When n = 250, rates of rejection due to poor fit were quite high. Overall, the results indicated that sample size played an important role with regard to model performance. In general, once *n* = 500, all models performed well under the conditions tested in this simulation.

CHAPTER 5

DEMONSTRATION OF THE SHIFTING INDICATORS MODEL WITH REAL DATA

In this chapter, I present a demonstration of the use of second-order LGMs with shifting indicators using real data. The purpose of this chapter is to exemplify how the model can be used in applied settings. This demonstration explicates the steps needed to run a second-order LGM with shifting indicators and compares the performance of the shifting indicators model to a model with the full set of items.

Methods

Participants

Data for this demonstration comes from the *Healthy Teens Longitudinal Study*, a study of adolescent social and academic development. A cohort of students in Georgia was assessed yearly from Grade 6 to 12. Students were randomly selected from nine middle schools located in six counties in Northeast Georgia. In sixth grade, 939 students were invited to participate and 745 accepted (79.3%).

The middle school data collection was a part of the Multisite Violence Prevention Project (MVPP, 2004, 2009). One scale is used in this demonstration. Responses from students who completed the scale at least once during grades 6-8 were included in the example dataset; the final sample comprised 720 students (48.2% girls; 48.9% White, 33.3% African American, 12.2% Hispanic, 1.7% Asian, 3.9% Multiracial/Other).

Measures

One scale from the Goals and Strategies measure, based on Hopmeyer and Asher (1997), was used in this demonstration. The Goals and Strategies measure presents respondents with vignettes of conflict scenarios, such as one student taking another student's seat. After a description of the scenario, a series of items assesses the respondent's 1) strategies for responding to the situation by asking "what would you do?" followed by a list of options and 2) goals in responding to the situation by asking "what would be your goal?" followed by a list of options.

In this demonstration, the scale that measures the goal of "maintaining a good relationship" was used. The scale is composed of four items, one for each of four vignettes. After each vignette, respondents rate how much they agree with the following statement about their goal in responding to the conflict if it had happened to them: "My goal would be trying to get along with this student." Five response categories were numbered from 1 to 5, and the endpoints were labeled *really disagree* and *really agree*. The middle response categories were not labeled with a verbal description. The data were subsequently recoded to range from 0 to 4.

Means, standard deviations, skewness, and kurtosis of the items at each time point are presented in Table 5.1. At each of the three waves of data collection, the scale demonstrated acceptable internal consistency (see Table 5.1). Table 5.2 presents the correlations among all items.

Table 5.1

Example Data Item Means, Standard Deviations, and Scale Internal Consistency

	Grade 6				Grade 7				Grade 8			
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*M*	*SD*	*Skewness*	*Kurtosis*	*M*	*SD*	*Skewness*	*Kurtosis*
1	2.66	1.468	-.691	-.916	2.40	1.416	-.396	-1.076	2.28	1.398	-.289	-1.116
2	2.69	1.475	-.743	-.879	2.58	1.415	-.589	-.893	2.45	1.402	-.455	-1.012
3	2.72	1.496	-.789	-.839	2.54	1.452	-.508	-1.084	2.42	1.450	-.417	-1.124
4	2.73	1.500	-.788	-.859	2.51	1.508	-.498	-1.184	2.48	1.455	-.491	-1.090
Internal consistency	0.807				0.767				0.808			

Table 5.2

Example Data Item Intercorrelations

		Grade 6				Grade 7				Grade 8			
		Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4
Grade 6	Item 1												
	Item 2	.570											
	Item 3	.536	.672										
	Item 4	.501	.631	.728									
Grade 7	Item 1	.274	.308	.370	.355								
	Item 2	.320	.285	.341	.352	.571							
	Item 3	.287	.296	.307	.381	.526	.670						
	Item 4	.241	.293	.296	.359	.531	.670	.743					
Grade 8	Item 1	.229	.272	.274	.330	.325	.348	.335	.373				
	Item 2	.171	.271	.231	.273	.303	.405	.339	.401	.551			
	Item 3	.231	.265	.273	.347	.379	.412	.428	.420	.561	.691		
	Item 4	.197	.223	.252	.298	.367	.446	.389	.442	.556	.644	.780	

All correlations are significant at the *p* < .05 level

Estimation Method

Basic descriptive statistics were obtained using SPSS 19. All SEM-based analyses were conducted using Mp*lus* 6.

For the purpose of reducing the complexity of the demonstration, the data were considered continuous. Although ordered categorical data are inherently not continuous, under certain circumstances it may be acceptable to treat categorical data as continuous (Finney & DiStefano, 2006, p. 298-299). Specifically, at least five response categories should be present, and responses should be approximately normally distributed or only moderately non-normal (i.e., skew < |2| and kurtosis < |7|; Finney & DiStefano, 2006, p. 299). An inspection of the response frequencies and associated histograms indicated that responses favored the *really agree* end of the scale. However, all of the scale points were used, and skewness and kurtosis values were relatively small. In particular, skewness values did not exceed |0.8|, and kurtosis values did not exceed |1.2|. Further, these values did not exceed the limit for moderate non-normality indicated by Finney and DiStefano (2006, p. 299). Given the guidance provided in Finney and DiStefano (2006), treating the data as continuous in this case appears justified and substantially simplifies the demonstration. Furthermore, the inherent non-normality of the data was taken into account in the selection of the estimator.

For all SEM analyses, a maximum likelihood estimator with robust standard errors (MLR) was employed to account for the presence of non-normality of the data. When data are non-normal, chi-square values may be inflated and standard errors underestimated under maximum likelihood (ML) estimation (Finney & DiStefano, 2006, p. 273). MLR produces a chi-square test statistic and estimates of standard errors that are adjusted for the level of non-normality present in the data. MLR is an extension of the Satorra-Bentler method (Satorra &

Bentler, 2001) that accommodates missing data using a full information maximum likelihood approach (FIML; Muthén, 2009). FIML estimates parameters and standard errors based on all available information.

The dataset was screened for outliers using the macro given in DeCarlo (1997), which identifies multivariate outliers based on the Mahalanobis distance for each observation. Eleven cases met the criteria for multivariate outliers at the .05 level of significance. These outliers were retained, as they appeared to reflect true variation in the sample, rather than miscoding or other errors in the dataset.

Analyses

Three phases of data analysis were conducted. First, a shifting indicators model was developed. Second, the items were screened for longitudinal measurement invariance. The third phase of data analysis entailed fitting the full and shifting indicators second-order LGMs to the data. The performance of the two models was compared with regard to parameter estimates and measures of model fit.

1) Development of the shifting indicators model.

The first phase of data analysis was developing the shifting indicators model. The shifting indicators model is a reduced form of the full model. Under both cases, a linear growth trajectory was specified for the three equally-spaced time points. The first item was designated as the scaling indicator, which sets the metric of the measurement model. The intercept and loading for item 1 at each time point were set to zero and one, respectively. Measurement errors of corresponding items at adjacent time points were allowed to co-vary. For the full model, the latent trait at each measurement occasion was measured by all four items. Under the shifting indicators model, a different item was dropped at each time point. A graphical representation of the full and the shifting indicators model is provided in Figure 5.1

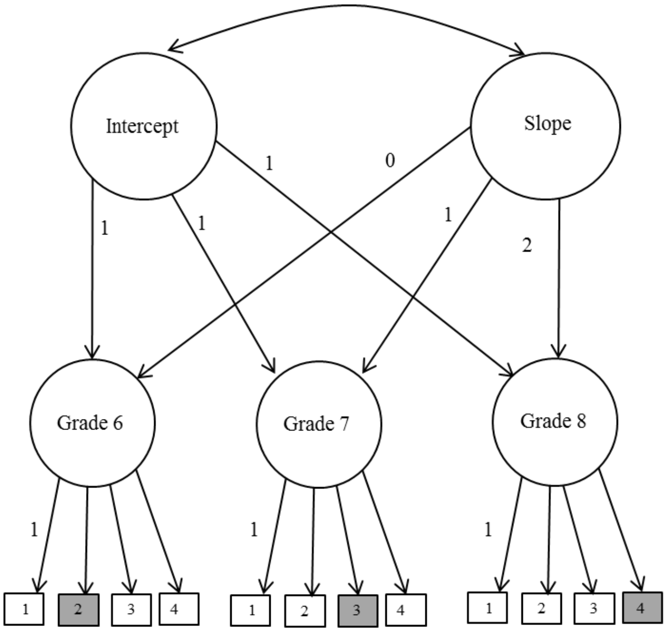


Figure 5.1. The full and shifting indicators models. The full model includes four items at each measurement occasion. In the shifting indicators model, one item per occasion is dropped, as shown in grey.

The purpose of this demonstration is to provide a step-by-step illustration of the application of a shifting indicators model and not to draw substantive conclusions about the development of the latent construct. It was beyond the scope of the demonstration to create a theoretically-based shifting indicators model. Instead, the development of the shifting indicators model was based on practical considerations. At each of the three time points, one item was dropped and three items were retained, in accordance with the three item per factor minimum suggested by Marsh, Hau, Balla, and Grayson (1998). At the first time point, item 2 was dropped. At the second time point, item 3 was dropped, and at the third time point, item 4 was dropped.

A primary consideration in developing a shifting indicators model is determining whether sufficient overlap in indicators exists to identify the model. As described earlier, Hancock and Buehl (2008) provided extensive guidance for determining whether the model will be identified. A series of matrices was developed based on this guidance, including the configuration matrix and the incidence matrix, presented in Figure 5.2.

Configuration Matrix

$$\text{Item}\;\begin{array}{c|ccc|} & \multicolumn{3}{c}{\text{Time}} \\ & 1 & 2 & 3 \\ \hline 1 & * & * & * \\ 2 & & * & * \\ 3 & * & & * \\ 4 & * & * & \end{array}$$

Incidence Matrix: Step 1

$$\text{Time}\;\begin{array}{c|ccc|} & \multicolumn{3}{c}{\text{Time}} \\ & 1 & 2 & 3 \\ \hline 1 & & & \\ 2 & 1 & & \\ 3 & 1 & 1 & \end{array}$$

Incidence Matrix: Step 2

$$\text{Time}\;\begin{array}{c|ccc|} & \multicolumn{3}{c}{\text{Time}} \\ & 1 & 2 & 3 \\ \hline 1 & & & \\ 2 & \cancel{1} & & \\ 3 & \cancel{1} & \cancel{1} & \end{array}$$

Figure 5.2. Matrices involved in determining sufficient overlap.

In the configuration matrix, asterisks identify the items that measure the latent construct at each time point. The incidence matrix identifies which time points share constrained items in common. In the present case, all three time points share constrained items in common, so the matrix contains all 1's. In the second step of the incidence matrix, vertical and horizontal lines are drawn through the non-zero elements of the incidence matrix. The shifting indicator model has sufficient overlap for identification if there are no zeros left that are not crossed out in the matrix. In this case, the incidence matrix had no zeros whatsoever, so the criteria for sufficient overlap for model identification was met.

2) Longitudinal measurement invariance.

As discussed earlier, it is important to obtain evidence that the measures exhibit adequate longitudinal measurement invariance. Ferrer et al. (2008) describe two methods for examining longitudinal measurement invariance of second-order LGMs. The first method involves running a series of nested confirmatory factor analytic (CFA) models where additional equality constraints are invoked at each step and models are compared using chi-square difference tests. The second method involves running a series of nested LGMs with equality constraints added at each step. Nested models are compared using a chi-square difference test. In this demonstration, the LGM-based method was utilized.

To assess configural invariance, the second-order LGMs were run with all loadings and intercepts freely estimated. In the second step, corresponding loadings were constrained to equality to test for metric invariance. In the third step, corresponding loadings and intercepts were constrained to equality to test for scalar invariance. At each step, the effect of the added equality constraints on the fit of the model was assessed with a chi-square difference test.

3) Comparison of second-order LGMs with identical and shifting indicators.

The third phase of data analysis involved running the full second-order LGM and the shifting indicators model. The models were then compared on model fit and estimates of growth parameters.

Assessing the fit of the models.

Model fit was determined using the chi-square fit index, the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). As described earlier, the chi-square fit index measures the discrepancy between the original and model-implied covariance matrices. A non-significant chi-square value indicates that the matrices are not significantly different. However, in large samples chi-square may be sensitive to trivial differences in the matrices. Therefore, this measure of model fit is often used in concert with other fit indices. In this demonstration, CFI, TLI, and RMSEA are also reported. As described earlier, values of CFI and TLI that exceed .95 are indicative of good fit (Hu & Bentler, 1998, 1999), as are values of RMSEA of less than .06 (Hu & Bentler, 1998).

Obtaining estimates of growth parameters.

The full model and shifting indicators model were compared on estimates of the five key growth parameters (i.e., slope mean, intercept mean, slope variance, intercept variance, and slope/intercept covariance) and their standard errors. Estimates were compared in terms of the statistical significance of parameter estimates, the direction of parameter estimates, and the magnitude of the parameter estimates and their standard errors.

Results

Examining Longitudinal Measurement Invariance

In order to test for measurement invariance, a series of nested LGMs was fit to the data and compared using the chi-square difference test. This process was conducted with both the full model and the shifting model. Table 5.3 displays the results of these analyses.

Table 5.3
Sequential Chi-square Difference Tests of Longitudinal Invariance

		Full		Shifting	
Test Type	Comparison	Chi-square difference	*df* difference	Chi-square difference	*df* difference
Metric Invariance	Model 1 vs 2	5.01	6	2.74	3
Scalar Invariance	Model 2 vs 3	12.01	6	5.10	3

Model 1: No invariance constraints
Model 2: Invariance of loadings
Model 3: Invariance of loadings and thresholds
Under MLR estimation, an adjustment to the chi-squared values and degrees of freedom used to conduct the difference test is required. All difference tests were adjusted according to the guidance provided on the Mplus website: www.statmodel.com.

Model 1 allowed intercepts and loadings to freely vary across time. This model was used to test for configural invariance, or that the factor structure is similar across time points. The overall fit of the model was acceptable for the full model (χ^2 (44) = 108.801, p < .001, CFI = .975, TLI = .962, RMSEA = .045) and the shifting indicators model (χ^2 (21) = 39.928, p = .008, CFI = .987, TLI = .978, RMSEA = .035). In both full and shifting cases, the chi-square value was significant, indicating that the model-implied covariance matrix was significantly different than the original covariance matrix. However, values of the other fit indices were indicative of good fit, and factor loadings were significant at each time point.

Model 2 constrained corresponding loadings to equality. A comparison of Model 1 and Model 2 provided evidence that metric invariance held for these data under the full and the shifting cases. A non-significant chi-square difference test indicated that adding equality constraints to corresponding loadings did not significantly decrease model fit.

Model 3 constrained corresponding factor loadings and item intercepts to equality. A comparison of Model 2 and Model 3 indicated that scalar invariance held for these data under the full and the shifting cases. A non-significant χ^2 difference test indicated that adding equality constraints to corresponding item intercepts did not significantly decrease model fit.

Comparison of Second-order Growth Model with Identical and Shifting Indicators

Because the data under both the full and the shifting cases exhibited longitudinal measurement invariance at the configural, metric, and scalar levels, the next step of data analysis was conducted. In this phase, the two models were compared on measures of model fit and estimates of growth parameter estimates.

Model fit.

Each of the models exhibited acceptable model fit, as displayed in Table 5.4. In both cases, the chi-square value was significant. While a significant chi-square value can be a sign of poor model fit, the chi-square test may be undesirably sensitive to trivial differences in the original and reproduced matrices under large sample sizes, as in this case. All other measures of fit yielded values indicative of good fit according to the recommendations outlined above.

Overall, the shifting indicators model appeared to exhibit better fit than the full model, although the differences in the fit indices were quite small. This result falls in line with the results of the simulation study, which found that a reduction in the number of indicators per factor led to better model fit according to the chi-square fit statistic, the CFI, and the TLI.

Table 5.4
Fit Indices of Second-order LGMs with Identical and Shifting Indicators

Model	Chi-square						
	Value	*df*	*p*-value	CFI	TLI	RMSEA	RMSEA CI
Full	127.042	56	<.0001	0.972	0.967	0.038	.032-.052
Shifting	47.79	27	0.0081	0.986	0.982	0.037	.017-.048

RMSEA CI: RMSEA confidence interval

Comparison of the growth parameter estimates.

Table 5.5 presents results of the comparison of growth parameters estimated by the full and shifting indicators models. For both models, the direction of the estimates was in concordance. That is, the slope mean and the intercept/slope covariance were negative, and all other values were positive. In all but one case, the statistical significance of the estimates was in agreement. In the full model, all five growth parameters were statistically significant at the .05

level. In the shifting indicators model, four growth parameters were statistically significant and one parameter, the intercept/slope covariance, was on the cusp of significance (i.e., .052).

Overall, the estimates and standard errors were quite similar across the two models. The largest difference in the parameter estimates occurred for the intercept variance, which was estimated as .535 in the full model and .573 in the shifting model. The largest difference in the standard errors was also observed in the intercept variance, which was estimated as .09 in the full model and .1 in the shifting model. Across the board, the standard errors were slightly larger in the shifting model. On the whole, these differences in the parameter estimates and standard errors seem trivial.

Table 5.5

Estimates of Key Parameters Obtained by Second-order LGMs

	Estimate	Standard Error	Est./S.E.	*p*-value
Intercept Mean				
Full	2.557*	0.046	55.179	<.001
Shift	2.574*	0.047	54.545	<.001
Slope Mean				
Full	-0.111*	0.023	-4.744	<.001
Shift	-0.132*	0.025	-5.171	<.001
Intercept Variance				
Full	0.535*	0.090	5.940	<.001
Shift	0.573*	0.100	5.734	<.001
Slope Variance				
Full	0.128*	0.040	3.162	0.002
Shift	0.134*	0.044	3.032	0.002
Intercept/Slope Covariance				
Full	-0.096*	0.046	-2.078	0.038
Shift	-0.103	0.053	-1.943	0.052

* *p* < .05

Figure 5.3 presents the model-implied growth trajectories for the full and the shifting indicators models. As displayed in the figure, the growth trajectory of the shifting indicators model had a slightly higher intercept and a slightly larger negative slope, resulting in an overall larger drop in the latent construct over the three time points in comparison to the full model. However, this difference is trivial. The overall model-implied decrease in the latent construct over time was .222 for the full model and .264 for the shifting model, for a total difference of .042 across a five-point scale.

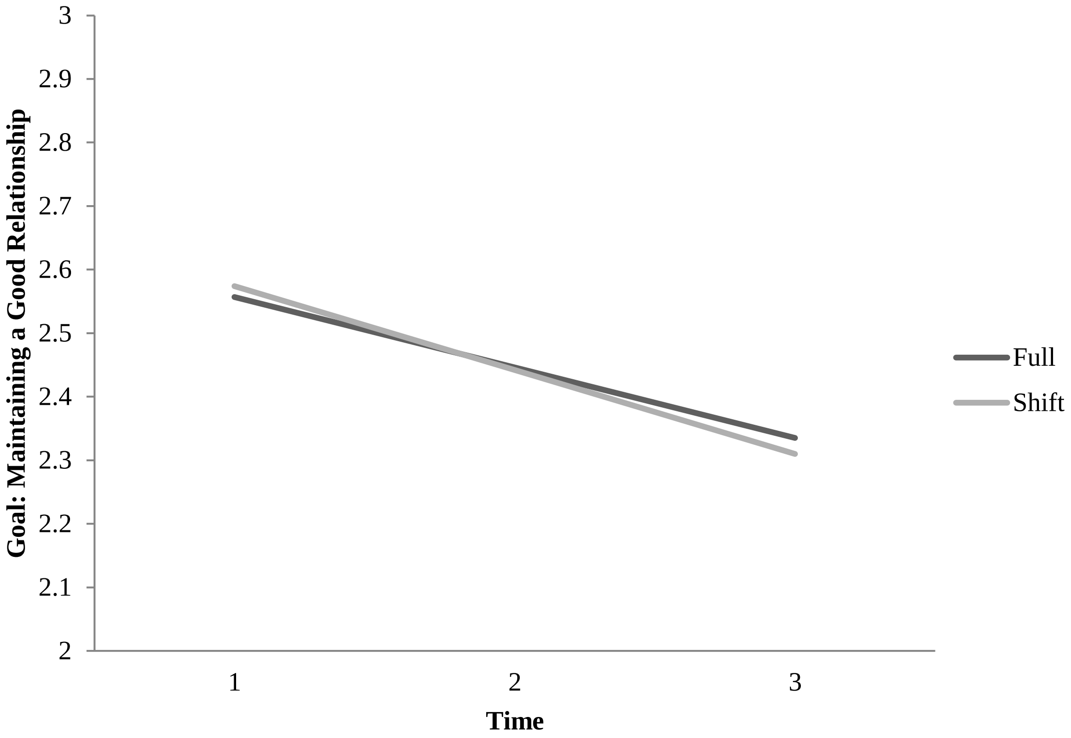


Figure 5.3. Growth trajectories implied by the full and shifting indicators model

In sum, in this demonstration the full and shifting indicators models produced similar results. The fit of the models, the direction of the parameter estimates, and the magnitude of the parameter estimates and standard errors were comparable. In every case, the significance of the parameter estimates was in concordance except for the intercept/slope covariance. While this parameter was significant in the full model, it was on the cusp of significance (i.e., .052) in the shifting model. The overall model-implied growth trajectories were virtually the same across the two models.

The results of this demonstration provide evidence that the shifting indicators model can produce similar results as a corresponding full model under real-world circumstances. The data employed in this demonstration had good model fit and exhibited longitudinal measurement invariance, but they were not contrived (e.g., missingness was encountered; data were Likert-type instead of continuous). Even under the current circumstance when items were dropped arbitrarily, the shifting indicators model produced similar estimates to the full model. Possible real-world applications of the shifting model include circumstances when items are a) omitted in a planned fashion to mirror developmental theory or b) lost due to unplanned circumstances. The results of this demonstration of the shifting indicators model are promising, even when items are lost without reference to theory.

CHAPTER 6

DISCUSSION

Longitudinal data analysis, as many developmental researchers know, can be messy and fraught with complication. Researchers may uncover problems with some items after data collection has begun, may have evolving ideas of how to best measure a construct, or may have a theoretical basis for wanting to measure a construct differently as participants mature. A benefit of the shifting indicators model is its flexibility in light of these types of circumstances. The shifting indicators model provides a mechanism for measuring growth that accommodates a changing array of items over time.

The purpose of this study was to investigate the performance of the shifting indicators model to provide information about the circumstances under which it may be appropriate. In the next section, I summarize the results of the study, make connections to what was previously known about the model, and explore the implications of these findings.

Conclusions and Implications

Convergence to a Proper Solution

The result of the simulation study indicated that non-convergence was not a problem in either the full model or the shifting indicators model. Only 14 of 76,000 datasets failed to reach convergence, for an overall convergence rate of 99.98%. Inadmissible solutions were also extremely rare (overall convergence rate: 99.67%) and were virtually non-existent when the sample size was greater than 250. For all cells, the rate of proper convergence was greater than 95%.

These results were unsurprising in light of previous research. The inverse relationship between sample size and frequency of inadmissible solutions is well documented in the SEM literature (e.g., Anderson & Gerbling, 1984; Boomsma, 1985; Chen, Bollen, Paxton, Curran, & Kirby, 2001). Studies by Leite (2007) and Hamilton, Gagné, and Hancock (2006) demonstrated this relationship in latent growth models.

Taken as a whole, these results are promising with regard to the shifting indicators model. Although rates of convergence to a proper solution were somewhat lower for the shifting indicators model, this effect was most pronounced when the sample size was 250. If rates of proper convergence were low, this would represent a significant drawback to the use of the shifting indicators model. However, the results of the simulation suggest that when shifting models are otherwise properly specified and an adequate sample size is employed, convergence is likely.

Bias in the Parameter Estimates and Standard Errors

Bias in the parameter estimates was not a substantial problem. When all solutions (i.e., admissible and inadmissible) were considered, the average values of relative bias of the growth parameters were acceptable across all cells. Replacing the nonconverged/inadmissible solutions did not have a substantial impact on the relative bias of the parameter estimates. For two cells, replacing the nonconverged/inadmissible solutions caused the relative bias of the slope variance to exceed the acceptable limit (i.e. .05). These cells had a sample size of 250, four items dropped at each of the three measurement occasions, and the lowest rates of proper convergence in the study.

Similar results were obtained with regard to the relative bias in the standard errors of the parameter estimates when all solutions were considered. In this case, the average values of the relative bias of the standard errors of the growth parameters were acceptable across all cells. Cells in the shifting indicator condition tended to exhibit slightly more bias than the cells from the full model condition, but this difference was small.

The overall lack of bias observed in parameter estimates and standard errors when all solutions were considered mirrors the results found in Leite (2007). In his study of second-order LGMs, the relative bias of the parameter estimates and their standard errors was acceptable under all conditions. Results with inadmissible solutions were not described in that paper but are presented in a dissertation based on the same study (Leite, 2005). These results indicate that once inadmissible solutions were removed, several of the parameters (i.e., slope variance, intercept variance, and slope intercept/covariance) and their standard errors exhibited bias above the cutoff. Bias occurred most often in the small to moderate sample size conditions (i.e. 100, 200, 500) and was mostly associated with the estimates of slope variance and intercept/slope covariance and their standard errors. The bias values reported by Leite (2005) were larger than those found in the current study, likely because he included conditions (e.g., reliability) that were not manipulated in this study.

The demonstration of the shifting indicators model in Chapter 5 provided an additional window into the performance of the models with regard to parameter and standard error estimation. Across all five growth parameters, the estimates obtained by the shifting indicators model were similar to the full model parameter estimates. The values of the standard errors were larger in the shifting models, and in one case, the larger standard error led to a difference in the statistical significance of the associated parameter estimate (i.e., the slope/intercept covariance).

Together, the results indicate that an otherwise properly specified shifting indicators model is unlikely to have substantial problems with bias in parameter estimates and standard errors, so long as the sample size is moderate (i.e., >250). However, standard errors may be slightly larger than those produced by a corresponding full model. For parameter estimates on the borderline of being significant, the larger standard errors may contribute to a higher probability of Type II errors. That is, significance testing may falsely indicate that a parameter is not significantly different than zero, when a significant effect should have been found.

Efficiency

Efficiency was investigated by examining the average standard errors of the growth parameter estimates. An examination of these values indicated that efficiency was slightly better for the full models than the shifting models, although the magnitude of the difference was small. Further, partial eta-squared values indicated that there were no practical differences in efficiency between the full and shifting indicators models. Several other conditions, most notably sample size, did have practically significant effects (i.e., partial eta-squared exceeded .05) on efficiency. As sample size increased, efficiency was improved for all five growth parameters. Additional practically significant effects of other conditions were obtained for the efficiency of selected growth parameters. In general, efficiency was reduced when larger numbers of items across more time points were either dropped or had low loadings. This is not surprising, because efficiency should be higher for situations in which estimates are based on more information.

These results are encouraging in terms of future applications of the shifting indicators model. Although the shifting indicators models were slightly less efficient, the difference was small. The differences in efficiency between the full and the shifting indicators models were more notable when the factor loading condition was .7. This condition was intended to simulate an unplanned loss in data. These results suggest that any reduction in efficiency will be more pronounced when items with large factor loadings are unexpectedly lost, as opposed to when items with low loadings are dropped because they are no longer developmentally appropriate.

Chi-square Fit Statistic

The performance of the chi square fit statistic generally behaved as expected given previous research findings. Model fit (as measured by Type I error rates) was worse when sample size was small. This result falls in line with Leite (2007), who found that for properly specified second-order LGMs, model fit increased as sample size increased.

In the current study, the number of items per factor appeared to play an important role in differentiating rates of rejection between the full and the shifting indicators models. For almost every case, the shifting indicators models, which had fewer indicators per factor, had lower Type I error rates than the corresponding full models. Within the shifting indicators models, the models with four dropped items tended to have fewer Type I errors than models with two dropped items. This pattern, in which better estimates of model fit were associated with models that had fewer items per factor, is similar to findings from Leite (2007) and Marsh, Hau, Balla, & Grayson (1998). Also in accord with these previous findings, the effect was more pronounced at smaller sample sizes.

The results of the applied demonstration also suggested that the fit of the shifting model was better than the fit of the full model. Although both model types had significant chi-square values, the chi-square/degrees of freedom ratio for the full model (127.042/56 = 2.268) was further from the ideal value of 1.00 than that of the shifting model (47.79/27 = 1.77). Because the items dropped from the shifting model were chosen arbitrarily, there are no theoretical reasons for the fit of the model to be improved. It seems likely that the estimates for fit of the shifting model are slightly improved due to having fewer items per factor.

These simulation and applied data results, taken together, indicate that model fit, as measured by chi-square Type I error rates, may be improved simply by dropping items from the model. This result raises a potential caution for the future use of the shifting indicators model. It would be considered poor practice to conduct arbitrary, pre- or post-hoc modifications to second-order LGMs for the sole purpose of enhancing model fit.

Other Fit Indices

The average values of the CFI, TLI, and RMSEA were indicative of good fit under all conditions. The percentage of rejected models according to the CFI and TLI cut-points were informative in differentiating across sample size conditions. For the *n* = 250 condition, the rates of rejection were quite high for the CFI (range: 19-43%) and the TLI (range: 18-43%). For all other sample size conditions, the rates of rejection were generally less than 5%, and when *n* = 1000, rejection rates were nearly zero. Average RMSEA values also indicated that fit improved as sample size increased, although the rejection rate was unaffected by sample size (i.e., the rejection rate was zero across all conditions). The association between sample size and model fit mirrors results reported by Leite (2007), who found that the percentage of rejected models decreased as sample size increased.

As with the chi-square results, the same pattern of lower rejection rates for the shifting versus the full model was observed with the CFI and TLI. For the full model, when increasing numbers of items with high loadings were replaced with items with low loadings, and CFI and TLI rejection rates generally increased. Conversely, as more items were dropped from the shifting models, rejection rates improved. It is likely that a similar explanation holds for the CFI/TLI results as with the chi-square results. That is, there is an improvement of fit when the number of items per factor decreases. Several other studies (Ding, Velicer, & Harlow, 1995;

Kenny & McCoach, 2003) have indicated that mean fit values of CFI and TLI decrease as the number of items per indicator increase. At larger sample sizes, this effect was minimized; this result was also observed in the current study.

Taken as a whole, the model fit results for the shifting indicators models were favorable. Given a sufficient sample size (i.e., > 250), the chi-square values and measures of fit obtained by the CFI, TLI, and RMSEA led to a high percentage of replications that met the criteria for acceptable fit. This result suggests that for shifting indicators models that are not otherwise misspecified (beyond the omitted items), model fit does not substantially degrade due to the exclusion of items, and in fact appears to somewhat improve. This slight improvement in model fit appears to be reflective solely of a reduction in the number of items per factor.

Limitations and Future Directions

This study aimed to investigate the performance of a type of model that has not received extensive attention in the literature. Because the shifting indicators model has not been comprehensively tested, the simulation was designed to test the model under somewhat idealistic conditions. The reasoning behind this decision is that if the model fails to perform well under idealistic conditions, it will almost certainly not perform well under real-world conditions.

This study examined the performance of second-order LGMs with shifting indicators with simulated data that were continuous, multivariate normal, and had no missing values. The factor loadings for the non-shifting indicators were all set to .7, and item intercepts were set to 0. In order to gain a more complete understanding of the performance of the shifting indicators model, additional investigations using more realistic data should be conducted. For example, categorical data, data that violate the assumption of normality, and data that include missing values should be considered in future studies. Further, the current study only used models that were properly specified apart from the omitted items. Studies that investigate the effects of different types of model misspecification on the performance of the shifting indicators model are warranted. For example, the performance of the shifting indicators model could be compared to full models under differing degrees of longitudinal measurement invariance with different types of items (invariant/non-invariant) dropped in the shifting models.

There are many other arrangements of shifting models that could be investigated in the future. For example, researchers could investigate the impact of having different numbers of items dropped at different time points (e.g., dropping two items at one measurement occasion and four at another), dropping the same number of items at different time points (e.g., dropping two items from time one versus dropping two items at time two), or adding items instead of dropping items.

Another possible line of research pertains to comparing the shifting indicators model to other methods of accounting for missingness in the case where items are inadvertently lost due to technical (or other data collection) problems. In the case where items are unintentionally omitted, the shifting indicators model can be used to exclude these items from the analysis. However, it may also be possible to code those items as missing and employ other missing data techniques. It would be of interest to investigate similarities and differences between the shifting indicators model and other methods of accounting for missingness.

In addition to research into the technical aspects of the shifting indicators model, applied researchers may benefit from additional work considering the practical aspects of deciding when to use the model. For example, the current study indicated that using the shifting indicators model may slightly increase bias (for at least some parameters and standard errors) and decrease efficiency. Some cautious researchers may be hesitant to consider dropping items for this reason, even if they are using some items that may be no longer developmentally appropriate. One aspect of the shifting indicators model that was not explored in this study relates to the experience of responding to items that are developmentally inappropriate. For example, if many items seem too childish to an adolescent population, it seems reasonable that at least some respondents may conclude that the survey is not worth taking seriously. These respondents could subsequently rush through the survey, be less thoughtful in their responses, or drop out of the study altogether. Longitudinal researchers need to consider the survey-taking experience as they seek to maintain the engagement of participants over time and avoid attrition. This type of concern cannot be investigated through simulation studies, but nonetheless represents an important consideration for applied researchers and an avenue for future research.

Guidance for Applied Researchers

The results of this study were generally promising in terms of the performance of the second-order LGM with shifting indicators. No serious problems were uncovered with regard to proper convergence, bias, efficiency, or model fit. However, as discussed earlier, the data used in this study were simulated under idealistic conditions. To the extent that real data differs from these conditions, the performance of the shifting indicators model may deteriorate. However, the demonstration in Chapter 5, in accord with Hancock and Buehl (2008), provided additional evidence that the shifting model is able to perform similarly to a corresponding full model using real data. In that demonstration, the parameter estimates, standard errors, and model implied growth trajectories were generally comparable (although in one case the significance test of a parameter estimate was not in agreement between the full and shifting models).

Researchers who wish to apply the shifting indicators model should consider both theoretical and methodological issues. For example, as discussed in Chapter 2, there should be a strong theoretical basis for identifying which items should be dropped. In terms of methodological considerations, the results of this study indicate that relatively large sample sizes (>250) are needed in order for the model to perform adequately. Somewhat larger sample sizes may be needed for the shifting indicators model to match the performance of full models. Furthermore, the results indicated that it may be unwise to have a large number of shifting items at multiple measurement occasions.

The overall findings from this study support the continued investigation into the performance of second-order LGMs with shifting indicators. Given the growing importance of longitudinal data methods and the flexibility of the shifting indicators model, this method of measuring growth may become an increasingly popular tool for researchers in the social sciences.

REFERENCES

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3-16.

Andersen J. C., & Gerbling, J. C. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices from maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155-172.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis, 95*(1), 1-22.

Bandalos, D. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.) *Structural equation modeling: A second course* (pp. 385-426). Greenwich, CT: Information Age.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong & M. H. M. Van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (Vol. 13, pp. 153-175). New York, NY: Oxford University Press.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL Maximum Likihood estimation. *Psychometrika, 50*, 229-242.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*, 13-20.

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*, 456-466.

Byrne B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321.

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.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199.

Chan, K.-Y., Drasgow, F., & Sawan L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*, 610-619.

Chen, F., Bollen, K., Paxton, P., Curran, P. J., & Colby, J. B. (2001). Improper solutions in structural equation models. *Sociological Methods in Research, 29,* 468-508.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 107–112.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional model on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2,* 292-307.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). The effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2*, 119-144.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development, 6*(2-3), 74-96.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495-515.

Enders, C. (2001). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352-370.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology (Gott), 4,* 22-36. doi: 10.1027/1614-2241.4.1.22.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.) *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *27,* 359-374.

Fischer, G. H. (1989). An IRT based model for dichotomous longitudinal data. *Psychometrika, 56,* 599-624.

Fitzpatrick, A. R. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice, 27*(4), 34-40.

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

Glück J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitation of four different approaches. *Methods of Psychological Research, 2*(1). Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/home.html

Hamburger, M. E., Basile, K. C., & Vivolo, A. M. (2011). *Measuring bullying, victimization, perpetration, and bystander experiences: A compendium of assessment tools.* Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Hamilton, J., Gagné, P., & Hancock, G. R. (2003, April). *The effect of sample size on latent growth models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hancock, G. R., & Buehl, M. M. (2008). Second-order latent growth models with shifting indicators. *Journal of Modern Applied Statistical Methods, 7*, 39-55.

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 470-489. doi: 10.1207/S15328007SEM0803_7

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluation longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 171-196). Greenwich, CT: Information Age Publishing.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York: Springer.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research, 26*, 329-364.

Hopmeyer, A., & Asher, S. R. (1997). *Children's responses to two types of peer conflict situations.* Symposium conducted at the biennial meeting of the Society for Research in Child Development, Washington, D.C.

Horn J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

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

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

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003) A comparison of linear, fixed common item, and concurrent parameter estimation procedures in capturing academic growth. *Journal of Experimental Education, 71,* 229-250.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136-153.

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,* 333-351.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Kolen M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York: Springer.

Lake, C. J., Gopalkrishnan, P., Sliter, M. T., Withrow, S. (2010). The Job Descriptive Index: Newly updated and available for download. *The Industrial-Organizational Psychologist*, *48,* 47-51.

Lance, C. E., Vandenberg, R. J., & Self, R. M. (2000). Latent growth models of individual change: The newcomer adjustment. *Organizational Behavior and Human Decision Processing, 83*, 100-140. doi: 10.1006/obhd.2000.2904

Leite, W. L. (2005). *A comparison of latent growth models for constructs measured by multiple items* (Doctoral dissertation). Retrieved from https://repositories1.lib.utexas.edu/2152/1609/leitew1123.pdf?sequence=2

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling, 14,* 581–610.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Marsh H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1*, 317-359.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181-220.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. E. Cattel (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561-614). New York, NY: Plenum.

McArdle, J. J., & Grimm, K. J. (2011). An empirical example of change analysis by linking longitudinal item response data from multiple tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 71-88). New York, NY: Springer.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Multisite Violence Prevention Project. (2004). The Multisite Violence Prevention Project: Background and overview. *American Journal of Preventive Medicine, 26*(1), 3-11.

Multisite Violence Prevention Project. (2009). The ecological effects of universal and selective violence prevention programs for middle school students: A randomized trial. *Journal of Clinical Psychology, 77*, 526-542.

Muthén, L. (2009, March 23). Re: MLR acronym? [Online forum comment]. Retrieved from http://www.statmodel.com/discussion/messages/11/2156.html?1317086403

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,* 599–620.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Pastor D. A. & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement, 30*, 100-120. doi: 10.1177/0146621605279761

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287-312.

Pettit, G. S., Keiley, M. K., Laird, R. D., Bates, J. E., Dodge, K. (2007). Predicting the developmental course of mother-reported monitoring across childhood and adolescence from early proactive parenting, child temperament, and parents' worries. *Journal of Family Psychology, 21*, 206-217. doi: 10.1037/0893-3200.21.2.206

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: The Danish Institute for Educational Research.

Reise, S. R., Widaman, K. F., & Pugh, R. H. (1993) Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

SAS Institute, Inc. (2011). *SAS/IML® 9.3 User's Guide*. Cary, NC: Author.

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

Sayer A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 177-200). Washington, DC: American Psychological Association.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

Singer J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Takane Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Tisak, J., & Meredith, W. (1990). Descriptive and associative development models. In A. Von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 2, pp. 387-406). Boston, MA: Academic Press.

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*(4), 4-70. doi: 10.1177/109442810031002

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76,* 318-336. doi: 10.1007/s11336-011-9202-z

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report 87-24). Princeton, NJ: Educational Testing Service.

APPENDIX A:

EXAMPLE FULL MODEL SYNTAX

Example of SAS Program to Generate Data and Analyze Models for a Full Model with All Loading Magnitudes of .7

```
/* STEP 1: Generate MPlus syntax files */

options noxwait xsync;
proc iml workspace = 90000;

%macro mplus;

%do n = 1 %to 4;
n = {250, 500, 750, 1000};
nn = n[&n,];
s = {62224 87136 41842 27611};
ss = s[&n];

file "C:\dis\model-7000-n&n..inp";

put ("MONTECARLO:");
put ("NAMES = y11-y18 y21-y28 y31-y38;");
put ("NOBSERVATIONS = ") @; put (nn) @; put (";");
put ("NREPS = 1000;");
put ("SEED = ") @; put (ss) @; put (";");
put ("REPSAVE = ALL;");
put ("SAVE = C:\dis\model-7000-") @;put ("n") @; put ("&n.") @; put
("*.DAT;") ;
put ("MODEL MONTECARLO:");
put (" t1 by y11-y18*.7;") ;
put (" t2 by y21-y28*.7;") ;
put (" t3 by y31-y38*.7;") ;
put (" y11-y38*.51;") ;
put (" i s | t1@0 t2@1 t3@2;");
put (" [i@1 s@1];");
put (" i*1 s*.2;");
put (" i with s*.179;");
put (" t1-t3@1;");
put (" [y11-y38@0];");

put (" MODEL:");
put (" t1 by y11-y18*.7;") ;
put (" t2 by y21-y28*.7;") ;
put (" t3 by y31-y38*.7;") ;
put (" y11-y38*.51;") ;
put (" i s | t1@0 t2@1 t3@2;");
put (" [i@1 s@1];");
put (" i*1 s*.2;");
put (" i with s*.179;");
put (" t1-t3@1;");
put (" [y11-y38@0];");
put ("OUTPUT: tech9;");

closefile "C:\dis\model-7000-n&n..inp";

%end;

/*Call Mplus and run files created above*/

%do n = 1 %to 4;

X call "C:\Program Files\Mplus\Mplus.exe" "C:\dis\model-7000-n&n..inp"
"C:\dis\model-7000-n&n..out";

%end;

%mend;
%mplus;

run;

/* STEP 2: Generate Mplus syntax files to run data */

options nodsnferr noxwait xsync cleanup nonotes NOSOURCE NOSOURCE2
errors=0;

proc iml workspace = 90000;

%macro analyze1;

%do n = 1 %to 4;
%do i = 1 %to 1000;

file "C:\dis\model-7000-run-n&n.i&i..inp" ;

put ("data: file = C:\dis\model-7000-") @; put ("n") @;
put ("&n.") @; put ("&i.") @; put (".dat;");
put ("variable: names = y11-y18 y21-y28 y31-y38;");
put ("model: t1 by y11-y18*.7 (1-8);");
put ("t2 by y21-y28*.7 (1-8);");
put ("t3 by y31-y38*.7 (1-8);");
put ("y11-y38*.51;");
put ("i s | t1@0 t2@1 t3@2;");
put ("[i*1 s*1];");
put ("i*1 s*.2;");
put ("i with s*.179;");
put ("t1-t3@1;");
put ("[y11-y38@0];");
put ("output: TECH1 MODINDICES (ALL 5);");
put ("savedata: results are C:\dis\model-7000-run") @; put ("n") @;
put ("&n.") @; put ("i") @; put ("&i.") @; put (".res;");

closefile "C:\dis\model-7000-run-n&n.i&i..inp" ;

%end;
%end;

/*call Mplus and run files created above*/

%do n = 1 %to 4;
%do i = 1 %to 1000;

X call "C:\Program Files\Mplus\Mplus.exe" "C:\dis\model-7000-run-
n&n.i&i..inp"
"C:\dis\model-7000-run-n&n.i&i..out";

%end;
%end;

%mend;
%analyze1;

/* STEP 3: Write data to .dat file*/

options nodsnferr noxwait xsync cleanup errors=1 notes;
proc iml workspace = 90000;

%macro writedata;

%do n = 1 %to 4;
%do i = 1 %to 1000;

x copy "C:\dis\null-full.txt" "C:\dis\7000.dat";
x copy "C:\dis\model-7000-runn&n.i&i..res" "C:\dis\7000.dat";

data one;
infile "C:\dis\7000.dat" truncover scanover flowover;
input im sm y11-y18 iv iws sv r11-r18 r21-r28 r31-r38 imse smse se11-
se18 ivse iwsse svse rse11-rse18 rse21-rse28 rse31-rse38 chi chidf
chip cfi tli logh0 logh1 freepar aic bic bica rmsea rmsealo rmseahi
rmseapr srmr;
rep= &i;
n = &n;
model = 7000;

file "C:\dis\results-7000.dat" mod;
put rep 4.0 +1 @; put n 1.0 +1 @; put model 4.0 +1 @;
put im 6.4 +1 @;
put sm 6.4 +1 @;
put iv 6.4 +1 @;
put iws 6.4 +1 @;
put sv 6.4 +1 @;
put imse 6.4 +1 @;
put smse 6.4 +1 @;
put ivse 6.4 +1 @;
put iwsse 6.4 +1 @;
put svse 6.4 +1 @;
put chi 9.4 +1 @;
put chidf 3.0 +1 @;
put chip 5.4 +1 @;
put cfi 6.4 +1 @;
put tli 6.4 +1 @;
put logh0 10.4 +1 @;
put logh1 10.4 +1 @;
put freepar 2.0 +1 @;
put aic 9.4 +1 @;
put bic 9.4 +1 @;
put bica 9.4 +1 @;
put rmsea 5.4 +1 @;
put rmsealo 5.4 +1 @;
put rmseahi 5.4 +1 @;
put rmseapr 5.4 +1 @;
put srmr 6.4 +1 ;

run;

%end;
%end;

%mend;
%writedata;
```

APPENDIX B:

EXAMPLE SHIFTING MODEL SYNTAX

Example of SAS Program to Analyze a Shifting Model with Two Items Dropped at Time 1 and Time 3

```
/* STEP 1: Generate Mplus syntax files to run data */

options nodsnferr noxwait xsync cleanup nonotes NOSOURCE NOSOURCE2
errors=0;

proc iml workspace = 90000;

%macro analyze1;

%do n = 1 %to 4;
%do i = 1 %to 1000;

file "C:\dis\model-7221-run-n&n.i&i..inp" ;

put ("data: file = C:\dis\model-7000-") @; put ("n") @;
put ("&n.") @; put ("&i.") @; put (".dat;");
put ("variable:");
put ("names = y11-y18 y21-y28 y31-y38;");
put ("usevariables = y13-y18 y21-y28 y31-y36;");
put ("model: t1 by y13-y18*.7 (3-8);");
put ("t2 by y21-y28*.7 (1-8);");
put ("t3 by y31-y36*.7 (1-6);");
put ("y13-y18*.51;");
put ("y21-y28*.51;");
put ("y31-y36*.51;");
put ("i s | t1@0 t2@1 t3@2;");
put ("[i*1 s*1];");
put ("i*1 s*.2;");
put ("i with s*.179;");
put ("t1-t3@1;");
put ("[y13-y18@0 y21-y28@0 y31-y36@0];");
put ("output: TECH1 MODINDICES (ALL 5);");
put ("savedata: results are C:\dis\model-7221-run") @; put ("n") @;
put ("&n.") @; put ("i") @; put ("&i.") @; put (".res;");

closefile "C:\dis\model-7221-run-n&n.i&i..inp" ;

%end;
%end;
```

```
%do n = 1 %to 4;
%do i = 1 %to 1000;

X call "C:\Program Files\Mplus\Mplus.exe" "C:\dis\model-7221-run-
n&n.i&i..inp"
"C:\dis\model-7221-run-n&n.i&i..out";

%end;
%end;

%mend;
%analyze1;

/* STEP 2: Write data to .dat file*/

options nodsnferr noxwait xsync cleanup errors=1 notes;
proc iml workspace = 90000;


%macro writedata;

%do n = 1 %to 4;
%do i = 1 %to 1000;

x copy "C:\dis\null-full.txt" "C:\dis\7221.dat";
x copy "C:\dis\model-7221-runn&n.i&i..res" "C:\dis\7221.dat";

data one;
infile "C:\dis\7221.dat" truncover scanover flowover;
input im sm y1-y8 iv iws sv r11-r16 r21-r28 r31-r36 imse smse se1-se8
ivse iwsse svse rse11-rse16 rse21-rse28 rse31-rse36 chi chidf chip cfi
tli logh0
logh1 freepar aic bic bica rmsea rmsealo rmseahi rmseapr srmr;
rep= &i;
n = &n;
model = 7221;

file "C:\dis\results-7221.dat" mod;
put rep 4.0 +1 @; put n 1.0 +1 @; put model 4.0 +1 @;
put im 6.4 +1 @;
put sm 6.4 +1 @;
put iv 6.4 +1 @;
put iws 6.4 +1 @;
put sv 6.4 +1 @;
put imse 6.4 +1 @;
put smse 6.4 +1 @;
put ivse 6.4 +1 @;
put iwsse 6.4 +1 @;
put svse 6.4 +1 @;
put chi 9.4 +1 @;
put chidf 3.0 +1 @;
put chip 5.4 +1 @;
put cfi 6.4 +1 @;
put tli 6.4 +1 @;
put logh0 10.4 +1 @;
put logh1 10.4 +1 @;
put freepar 2.0 +1 @;
put aic 9.4 +1 @;
put bic 9.4 +1 @;
put bica 9.4 +1 @;
put rmsea 5.4 +1 @;
put rmsealo 5.4 +1 @;
put rmseahi 5.4 +1 @;
put rmseapr 5.4 +1 @;
put srmr 6.4 +1 ;

run;

%end;
%end;

%mend;
%writedata;
```