A LONGITUDINAL ITEM RESPONSE THEORY-LATENT GROWTH MODELING FOR MEASURING CHANGE

by

MEI LING ONG

(Under the Direction of Allan S. Cohen and Seock-Ho Kim)

ABSTRACT

Latent growth modeling (LGM) is a method commonly used for analyzing longitudinal data, derived from confirmatory factor analysis models as a special case of structural equation modeling. However, it has several limitations, such as the inability to assess measurement invariance in a longitudinal study. This study develops a longitudinal item response theory-latent growth modeling (LIRT-LGM) model, which can be viewed as a combination of a LGM model and an IRT model for the purpose of investigating growth or change in the latent variable(s). To motivate this study, an illustrative example was provided comparing the performance of the LGM and LIRT-LGM in analyzing depressive symptoms. The LIRT-LGM was used to analyze the data with both the one-parameter logistic (1PL) and the two-parameter logistic (2PL) models. A simulation study was presented to provide more detailed information about the performance of the LIRT-LGM. Test lengths, sample sizes, and effect sizes were manipulated in the simulation study. Type I error and power were compared for the LIRT-LGM to the LGM models. An analysis of a real data set from a measure of depressive symptoms indicated the performances of the LGM and LIRT-LGM were not consistent. For empirical results, the mean and variance of the slope of the LGM were statistically significant, indicating that depressive symptoms

increased and individual differences increased over three time points. On the other hand, the mean of the slope of the LIRT-LGM was not significant, but the variance of the slope of the 2PL version was significant. Results of the simulation indicated that the Type I error was controlled for most conditions. When the effect size was .3 with a sample size of 100 at $\alpha = .05$, the power was greater than .8. The results further showed that, when sample sizes, effect sizes, and test lengths increased, the performance of LIRT-LGM model was better than the LGM model.

INDEX WORDS: latent growth modeling, item response theory, longitudinal item response theory, maximum likelihood estimation

A LONGITUDINAL ITEM RESPONSE THEORY-LATENT GROWTH MODELING FOR MEASURING CHANGE

by

MEI LING ONG

B.A, Fu-Jen Catholic University, Taiwan, 1999

M.A, The University of Georgia, 2012

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

2017

A LONGITUDINAL ITEM RESPONSE THEORY-LATENT GROWTH MODELING FOR MEASURING CHANGE

by

MEI LING ONG

Major Professor:	Allan S. Cohen
	Seock-Ho Kim
Committee:	Gary J. Lautenschlager
	Zhenqiu (Laura) Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

To my family, my husband, and to dear Professors Allan S. Cohen and Seock-Ho Kim.

ACKNOWLEDGEMENTS

I sincerely appreciate those who supported and encouraged me throughout this process. I would like to thank my major professors, Drs. Seock-Ho Kim and Allan S. Cohen, for their guidance and technical support throughout this study, without which I would not have completed this dissertation. Words cannot express my deepest appreciation for all that both of you have done for me. In addition, I would like to thank the members of my committee, Drs. Gary Lautenschlager and Zenqiu (Laura) Lu, for their comments and helpful suggestions while completing this dissertation. Furthermore, I want to thank my friends, Yu Bao and Stephanie Short, who provided their opinions in terms of this dissertation.

Lastly and importantly, I wish to express my deepest appreciation to my parents, my elder brother and my younger sister for their support and encouragement. To my lovely husband, Man Kit, thanks for cooking lunch and dinner for me while I was researching, writing, and revising this study. Because of your unending encouragement and full support, I have had an opportunity to obtain my Dortoral degree.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Researchers may be interested in understanding why and how specific conditions and events, such as depression or growth in math aptitude, change over time. What is clear, however, is that these phenomena do not necessarily change at the same rate. For example, math ability generally increases steadily from elementary school to high school but not at the same rate or in the same way for all students. Researchers are typically interested in the overall patterns of change and whether the trend of the growth or decline is linear or some other pattern. They are also interested in investigating the different change processes across individuals.

One approach to studying these kinds of events is to include a time factor in the study's design. Cross sectional designs are often used to infer growth or change, but cross-sectional study designs do not account for within subject time. This is because cross-sectional studies collect data at one point in time. As a result, this kind of study cannot provide clear information related to individual change. Longitudinal research methods, on the other hand, do allow for observation of within subject change. As a result, these methods are often used to measure change over time, and longitudinal data analysis has become more accessible for use in accounting for the problems of growth and change (Singer & Willett, 2003).

1.1 Longitudinal Data Analysis

Longitudinal data indicate the "repeated time-ordered observation of an individual or individuals with the goal of identifying processes and causes of intraindividual change and of

interindividual patterns of intraindividual change in behavioral development" (Baltes & Nesselroade, 1979, p. 7).The longitudinal research method is characterized by the fact that there are repeated measures over time within individuals. The goal of longitudinal data analysis is to investigate changes in latent means over time and changes in individual differences over time regarding one or more outcome variables (Marsh & Grayson, 1994). There are two types of individual difference changes over time (Singer & Willett, 2003): (1) intra- (within-) individual change describes each person's individual growth trajectory; and (2) inter- (between-) individual differences change focus on whether different individuals present similar or different patterns of within-individual change.

Researchers apply longitudinal data to different topics, such as depression, for several reasons (Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006). First, cross-sectional data do not suffice for answering certain questions, such as causal relationships or predicted outcomes. Most research questions or theoretical assumptions are potentially caused by a change in causal relationships. Thus, when researchers want to obtain the best results related to the issue of change, they can collect and analyze longitudinal data. Second, in the past, when researchers wanted to investigate the cause-effect relationship between variables, they could only use true experimental design methods. After strictly controlling for confounding variables, researchers manipulated independent variables and observed the change in the dependent variables. Based on the results, researchers inferred whether a causal relationship existed between independent variables and dependent variables. However, most behavioral and social science (e.g., psychology or sociology) studies use survey methods to collect data from observations that are not manipulated. Thus, causal relationships could not be explored. Recent developments in longitudinal analyses, however, have enabled investigation of causal relationships by

incorporating a time factor in the analysis. In this way, the time factor is able to model the assumption of a path relationship.

1.2 Statement of Problem

Longitudinal data become an important ingredient in research to investigate developmental changes because they allow a researcher to investigate changes over time within individuals and differences between individuals at a baseline (Hedeker & Gibbons, 2006; McArdle & Bell, 2000). Information is normally collected across time points, such as a pretest and posttest, using two or more time points to infer a causal relationship. To make causal inferences, a study has to meet three conditions (Duncan et al., 2006). The first condition is covariation. This assumes that cause and effect are significantly correlated. The second condition is the temporal order of events. This assumes that cause precedes effect in time. The third condition is non-spuriousness. That is, other external factors that can influence the explanation of the dependent variables can be ruled out or can be controlled.

Traditionally, longitudinal data are analyzed using statistics such as a paired *t*-test, repeated measures ANOVA or ANCOVA, or auto-regressive models. However, these statistical techniques suffer from several limitations (Cho, Cohen, & Bottge, 2013; Maxwell & Tiberio, 2007). First, in the context of a pre-test and post-test analysis, a paired sample *t*-test is the simplest type of longitudinal analysis. It can be used to determine whether there is a significant mean difference between a pre- and post-test. Thus, this model is limited to two-time points: pre-test (Time 1) and post-test (Time 2). Second, repeated measures ANOVA or ANCOVA can be used to test the effects of a continuous dependent variable measured at several time points. Although these models are useful in understanding mean differences across time, there are some

limitations: (i) these models only work with balanced data; (ii) they cannot deal with missing data; (iii) these models ignore time ordering, which means that it is difficult to incorporate time-varying predictors; (iv) repeated measures ANOVA or ANCOVA are based on the strong assumption of sphericity, which means that the variances of every measure should be the same. Cho, Cohen, and Bottge (2013) indicated that it is possible to violate the assumption of sphericity in instructional intervention studies because the floor effects are likely to occur before the treatment, and (v) the repeated measures ANOVA or ANCOVA estimate average change in scores over time rather than individual differences in change. Third, auto-regressive models cannot provide an adequate generalization for more than two-time points (Duncan et al., 2006). Fourth, the dependent variable in traditional techniques is a sum score. In other words, they ignore item properties and test information. Thus, these methods do not work well for interpreting growth at the individual level.

Researchers should carefully consider the analytic method that best suits their studies. An appropriate analytic method must be compatible with both the hypotheses being tested and the data. These two elements provide evidence to support or refute the hypotheses of researchers. Unfortunately, traditional statistical methods tend to operate at the group level but fail to address hypotheses about the nature and the causes of change at the level of an individual (Hancock, Harring, & Lawrence, 2013). Most standard statistical approaches do not reflect the passage of time when examining growth. In order to deal with the limitations of these methods, latent growth modeling (LGM, Preacher, Wichman, MacCallum, & Briggs, 2008) has been proposed. LGM is also known as latent growth curve modeling (Preacher et al., 2008) or latent curve analysis (Meredith & Tisak, 1990) and has emerged from the field of structural equation modeling (SEM) (Duncan et al., 2006; Hancock et al., 2013; McArdle & Epstein, 1987).

LGM is a model for examining specific longitudinal data. It employs a concept derived from confirmatory factor analysis (CFA) models as a special case of SEM (Preacher et al., 2008; McArdle & Bell, 2000). A LGM model is commonly used for analyzing change over time in the behavioral and social sciences (e.g., education, psychology, or sociology). It can describe individuals' behavior in terms of initial levels and its growth trajectories to and from either linear or quadratic patterns. Additionally, the LGM model can help to define the change across individuals in both intercept and slope, and it simultaneously focuses on changes in covariances, variances, and mean values over time. Thus, the LGM procedure is unique because it combines individual and group levels of analysis (Duncan et al., 2006; Hancock et al., 2013). In the LGM model, the relationship between latent variables and indicators is similar to that of the CFA model. The effects of latent variables on their indicators are called factor loadings, which describe trends over time in variables that are repeated measures of the same observed variable.

Although the LGM model has several advantages, such as describing the change in the latent construct variable, it does not take fully model the measurement error of the SEM. The major drawback of LGM is that it lacks a mechanism for assessing measurement invariance in longitudinal research. Thus, LGM may not be the most useful method for drawing correct inferences related to change.

To this end, LGM has been extended to incorporate multiple indicators, namely second-order latent growth modeling, which can account for measurement invariance. A useful approach for simultaneously modeling measurement invariance, item properties, and test or item information in growth, is to consider incorporating IRT into the model. Thus, in this dissertation, we develop a longitudinal item response theory-latent growth modeling (LIRT-LGM). This model integrates LGM and IRT for use in analyzing longitudinal data to investigate change in

the latent variable. This dissertation presents a method of estimating model parameters with a maximum likelihood solution and a simulation study to investigate model performance under practical testing conditions.

Although the theory underlying the LGM and LIRT approaches have been available for some time, the LIRT-LGM is a new model. Thus, in the example, the LIRT-LGM model will be used to investigate changes in depressive symptoms in African-American adolescent girls. This will serve to motivate the simulation study.

1.3 The Purpose of the Study

Previous studies (e.g., Geiser, Keller, & Lockhart, 2013; Fleming at al., 2008) used LGM to analyze the depression scale. The observed variable for the LGM is computed as the sum of item scores. However, Fried and Nesse's (2015) study suggests that depression sum-scores do not add up because each item in the depression scale has its specific depressive symptoms. Thus, LGM may not be a useful approach for analyzing the depression scale. The first purpose of this study was to evaluate the performance of the LGM and LIRT-LGM for analyzing depressive symptoms. Two dichotomous IRT models will be implemented in the LIRT-LGM, the 1PL and the 2PL models. In the simulation study, we will provide more detailed information about the performance of the LIRT-LGM with attention paid to a number of items, sample sizes, and effect sizes. Sample sizes are important because appropriate sample sizes are needed to achieve sufficient power for statistical tests of interest (Hancock & French, 2013). In addition, this study considers that lengths of items are important because more items in a test generally will provide a greater amount of information (Baker, 2001). Further, effect sizes are also important as when effect sizes increase, so does power.

1. 4 Significance of the Study

The LIRT-LGM includes a latent variable based on observed (i.e., manifest) multiple indicators. As noted above, the 1PL and 2PL IRT models are used in this dissertation. In addition, the LIRT-LGM enables an estimate of a growth pattern for these latent variables. The difference between the LGM and LIRT-LGM is that the LGM estimates change based on a single observed variable at each time point, whereas the LIRT-LGM estimates change based on multiple indicators which are intended to estimate the latent constructs at each time point. The major advantage of the LIRT-LGM is that it allows researchers to simultaneously test measurement invariance and item properties across measurement time points. If measurement invariance holds, researchers can be more confident that the same latent constructs are measured at each time point. In addition, the LIRT-LGM provides information associated with the measurement characteristics of the indicators. The LIRT-LGM should have greater statistical power because it directly models the measurement structure of the indicators. Thus, this new model can simultaneously estimate individual differences in stability and change.

1.5 Organization of the Study

This dissertation describes the development of the LIRT-LGM model using two separate perspectives on modeling longitudinal item response data using the latent growth modeling method. Since the computation of observed variables in the traditional LGM and the LIRT-LGM models can be based on either a CTT- or IRT-based score, the first and second sections of Chapter 2 provide an overview of CTT and IRT. The third section describes some useful longitudinal item response theory models. The fourth section describes the basic concept of latent growth modeling that includes the structural equation modeling and IRT-SEM models.

The fifth section describes longitudinal item response theory-latent growth modeling (LIRT-LGM) as a combination of latent growth modeling and the IRT model with longitudinal data. The last section describes the research questions and rationale.

Chapter 3 presents the methodology of the study. The LGM and LIRT-LGM models are presented first with an empirical study and next with a simulation study. These two studies are used to evaluate the performance of the LIRT-LGM model. Chapter 4 presents the results of the empirical and simulation studies. In addition, the results of the LIRT-LGM model were compared with results from the traditional LGM analysis. Chapter 5 includes a discussion and conclusion that summarizes the methods and results and describes the practical importance and the significance of the study, limitations, and possible future studies.

CHAPTER 2

THEORETICAL BACKGROUND

The longitudinal item response theory-latent growth modeling (LIRT-LGM) that integrates LGM and IRT with longitudinal data is a new model. Thus, the purpose of this study is to compare the performance of both the LGM and LIRT-LGM and to examine the performance of the LIRT-LGM using empirical data and simulated data.

Since developmental research focuses on a latent construct, such as depression, that cannot be directly or simply observed, researchers need to use a set of items to build this latent construct. These items are assumed to be valid indicators of the construct, and scores that are derived from these items or from the scale are assumed to provide useful information associated with the level of a subject on that construct. Thus, these scores are treated as the representation of the latent construct (Edwards & Wirth, 2009). Classical test theory (CTT) and item response theory (IRT) are two of the techniques that can be used for scale construction.

In this section, we provide overviews of CTT, IRT, longitudinal item response theory (LIRT), and latent growth modeling (LGM) including structural equation modeling (SEM) and IRT-SEM. We then present a new model, longitudinal item response theory-latent growth modeling (LIRT-LGM). This study argues that LIRT-LGM can be a useful approach for measuring the change in latent constructs such as depression and may allow researchers to employ a more sophisticated framework for assessing change in the behavioral and social sciences.

2.1 The Basic Concept of Classical Test Theory (CTT)

The CTT is the earliest test theory for addressing measurement problems, such as test development. Charles Spearman introduced the concept of the observed score in 1904. Spearman (1904) argued that the observed score is comprised of the true- and an error- score. The observed score is the only manifest element, and the true- and error- score are latent. The observed score provides useful information, which can be used to improve the reliability of tests (Alagumalai & Curtis, 2005). Because the CTT mainly estimated the reliability of the observed scores in a test, it is also called classical reliability theory. That is to say, the CTT tried to estimate the strength of the relationship between the observed and true scores (Suen, 1990). The CTT is sometimes regarded as the true score theory because its theoretical root is based on the true score model, which assumes that the observed score (or total test score) is influenced by a true score and random error (Hammond, 2006). There are several assumptions in the CTT (Embretson & Reise, 2000; Hambleton & Jones, 1993). First, the errors are random and unrelated to true and observed scores. Second, true and error scores are uncorrelated. Third, the expected value of error scores in the population of respondents is zero. Fourth, the error scores on parallel tests are uncorrelated. The concept of CTT is more simply based on the total test scores, reliability, item difficulty, and item discrimination (Weiss & Yoes, 1991).

2.1.1 True Score Model

The central idea of the CTT can be defined as:

$$\text{Observed Score} = \text{True score} + \text{Error}. \quad (1)$$

Equation 1 defines that the observed score, which is the score individuals can obtain on the measuring instrument, is made up of two components, the “true score” and an “error score”

(Hammond, 2006; Wainer & Thissen, 2001; Drolet & Morrison, 2001, Hambleton & Jones, 1993; Weiss & Yoes, 1991). The true score, which is denoted as "the expected value of observed performance on the test of interest," (Hambleton, Swaminathan, & Rogers, 1991, p. 2) expresses the concept of ability in CTT. The true score is the average score if a respondent takes parallel forms of the test many times (Weiss & Yoes, 1991). The true score cannot be directly observed; thus, it is latent and inferred from the observed score. The term error, which is defined to be unsystematic or random and uncorrelated with the true score, describes the difference between a respondent's observed score and his or her true score (Crocker & Algina, 2008; Wainer & Thisesen, 2001). Equation 1 can be expressed in the usual notation as:

$$X = T + E, \tag{2}$$

where X is the observed score (or test score), T is the true score, and E is the error score. The path diagram of Equation 2 showed in Figure 1.

To discuss the mathematical terms, the equation for a respondent j can be represented as:

$$X_j = T_j + E_j, \tag{3}$$

where X_j is the observed score for respondent j, T_j is the true score for respondent j, and E_j is the

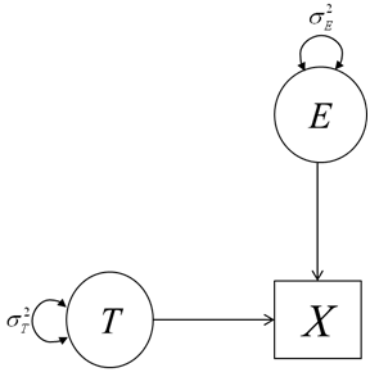

Figure 1. CTT model path diagram.

error for respondent *j*. The CTT assumes that each respondent has a true score. If the expected value of X_j is T_j, the expected value of E_j is zero (Wainer & Thissen, 2001; Lord, 1980). Thus, the equation of true score for respondent *j* is given as:

$$E(X_j) = T_j, \tag{4}$$

where $E(X_j)$ indicates the expected value of the random variable *X* for respondent *j*.

The error for respondent *j* from the Equation 3 is given as:

$$E_j = X_j - T_j, \tag{5}$$

where E_j and X_j are random variables, and T_j is a constant. The expected value of error for respondent *j* is:

$$\begin{aligned} E(E_j) &= E(X_j - T_j) \\ &= E(X_j) - E(T_j). \end{aligned} \tag{6}$$

Since $E(X_j) = T_j$ from Equation 4 and the expected value of T_j is constant, the expected value of error for respondent *j* can be written as (Wainer & Thissen, 2001):

$$E(E_j) = T_j - T_j = 0. \tag{7}$$

The CTT assumes that the scores between respondents are independent. When *T* and *E* are independent, the observed-score (or total score) variance (σ_X^2) can be decomposed into true score variance (σ_T^2) and error score variance (σ_E^2) (Lord & Novick, 1968). From Figure 1, this relation can be computed as:

$$\begin{aligned} \text{Var}(X) &= \text{Var}(T + E) \\ &= \text{Var}(T) + \text{Var}(E) + 2\text{Cov}(TE), \end{aligned} \tag{8}$$

where Cov(*TE*) is equal to zero. Thus, the equation of score variance can be written as:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{9}$$

According to Equation 9, measuring a respondent's true score is most reliable when the error variance of a test is small (Wainer & Thissen, 2001).

2.1.2 The Total Test Score

The total (or the number-correct) test score, which is the estimate of the ability of an individual, is defined as the sum of the item responses (scored 0, 1) for each respondent *j*. The equation of the total test score becomes

$$x = \sum_{i=1}^{n} k_{ji} \,, \tag{10}$$

where x is the total test score and k_{ij} are the item responses for respondent j and $i = 1, \ldots, n$ (Wainer & Thissen, 2001). Because the total test score is computed by summing the item responses, this will influence the reliability of the total test score.

2.1.3 Reliability

Reliability, which is one of the main concepts of the CTT, refers to the stability and consistency of assessment results, and it is the strength of the relationship between the observed and true scores. Reliability can be expressed as the Pearson's correlation (r) between the observed (X) and true (T) scores, and this correlation is known as the reliability index, ρ_{XT} (Crocker & Algina, 2008; Suen, 1990).

In general, the r between X and Y is given as (Suen, 1990):

$$\rho_{XY} = \frac{\sigma_{XY}}{(\sigma_X)(\sigma_Y)}, \tag{11}$$

where σ_{XY} is the covariance between *X* and *Y*, σ_X is the standard deviation of *X*, and σ_Y is the standard deviation of *Y*.

As stated above, when *T* and *E* are independent, ρ_{XT}, which also can be stated as the ratio of the standard deviation of true scores to the standard deviation of the observed scores (Lord & Novick, 1968), can be expressed as (Wainer & Thissen, 2001; Suen, 1990):

$$\rho_{XT} = \frac{\sigma_{XT}}{(\sigma_X)(\sigma_T)} \tag{12}$$

where ρ_{XT} is the correlation between the observed and the true scores, σ_T is the standard deviation of the true score, and σ_X is the standard deviation of the observed score, and σ_{XT} that is the covariance between *X* and *T* is given as:

$$\sigma_{XT} = \sigma_{(T+E)T} = \sigma_T^2 + \sigma_{TE} = \sigma_T^2. \tag{13}$$

where $X = T + E$ from Equation 2, and assuming that the true score and error score are independent suggests that $\sigma_{TE} = 0$. Equation 12 can be rewritten as:

$$\rho_{XT} = \frac{\sigma_T^2}{(\sigma_X)(\sigma_T)} \tag{14}$$

where σ_T^2 is the true score variance. When the ρ_{XT} relationship is strong as designated by a high *r*, *X* is better for reflecting *T*, and *X* can be viewed as a linear transformation of *T* (Suen, 1990). However, because *T* is unknown, we cannot directly estimate ρ_{XT} from the observed data. Thus, we can estimate the square of the reliability index (ρ_{XT}^2). The ρ_{XT}^2 as expressed in Equation 14 can be written as (Wainer & Thissen, 2001; Suen, 1990):

$$\rho_{XT}^2 = \frac{\left(\sigma_T^2\right)^2}{\left(\sigma_X^2\right)\left(\sigma_T^2\right)}$$

$$= \frac{\sigma_T^2}{\sigma_X^2}$$

$$= \frac{\sigma_{X-E}^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}$$

$$= 1 - \frac{\sigma_E^2}{\sigma_X^2}, \tag{15}$$

where ρ_{XT}^2 is referred to as the *reliability coefficient*, which can be stated as the correlation between scores on the parallel test forms (Crocker & Algina, 2008). For the parallel test forms, there are two forms, form *X* and form *Y*, of a test having scores *x* and *y*, respectively. If $E(x) = E(y) = t$ and $\sigma_x = \sigma_y$, two forms of a test are parallel. The correlation between the two parallel forms of observed scores will be yielded by:

$$\rho_{xy} = \frac{\sigma_{xy}}{(\sigma_x)(\sigma_y)}$$

$$= \frac{\sigma_{(t+e_x)(t+e_y)}}{(\sigma_x)(\sigma_y)}$$

$$= \frac{\sigma_{tt} + \sigma_{tey} + \sigma_{tex} + \sigma_{e_x e_y}}{\sigma_x \sigma_y}. \tag{16}$$

The last three terms of the numerator in Equation 16 will become zero because of the assumption of independence. Thus, Equation 16 will be yielded by:

$$\rho_{xy} = \frac{\sigma_{tt}}{\sigma_x \sigma_y} = \frac{\sigma_t^2}{\sigma_x^2}, \tag{17}$$

where $\sigma_{tt} = \sigma_t^2$ and $\sigma_x = \sigma_y = \sigma_x^2$. Equation 17 can be written as:

$$\rho_{xy} = \frac{\sigma_t^2}{\sigma_x^2}$$

$$= \frac{\sigma_{x-e}^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}$$

$$= 1 - \frac{\sigma_e^2}{\sigma_x^2}. \quad (18)$$

ρ_{xy} is the same as reliability coefficient ρ_{XT}^2. Even though ρ_{XT} cannot be directly estimated from the observed data, ρ_{XT}^2 can be estimated directly (Wainer & Thissen, 2001; Suen, 1990).

Reliability is a concept, and researchers apply reliability to designate the proportion of true score variance in a group's observed test scores. Since reliability is known, error variance, σ_e^2, is possible to estimate. An expression for σ_e^2 can be derived by using Equation 9. Dividing both sides by σ_x^2, the equation can be rewritten as:

$$1 = \frac{\sigma_t^2}{\sigma_x^2} + \frac{\sigma_e^2}{\sigma_x^2}, \quad (19)$$

since $\frac{\sigma_t^2}{\sigma_x^2} = \rho_{xy}$ from Equation 17, Equation 19 becomes

$$1 = \rho_{xy} + \frac{\sigma_e^2}{\sigma_x^2}$$

$$1 - \rho_{xy} = \frac{\sigma_e^2}{\sigma_x^2}$$

$$\sigma_x^2(1 - \rho_{xy}) = \sigma_e^2. \quad (20)$$

Thus, the standard error of measurement, which is the square-root of the error variance, can be given as:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xy}}. \quad (21)$$

The standard error of measurement can be used to build confidence intervals around the observed scores.

2.1.4 Item Difficulty and Item Discrimination

The CTT can provide useful information to guide the selection from item banks in order to improve the reliability of the total test score. The characteristics of an item difficulty index and item discrimination index can be related to the reliability of total test score, even though CTT is mainly focused on test-level information. Item difficulty under the CTT is described as the probability of respondents who provide a correct response to an item. This probability is referred to as the *p*-value of that item known as an item difficulty index. A high *p*-value indicates an easy item, whereas a low *p*-value indicates difficult items. Item discrimination is defined as the correlation between scores on a dichotomous item and on the total test, using the point-biserial correlation coefficient, which is the Pearson correlation between the dichotomous item variable and the continuous total score variables (Alagumalai & Curtis, 2005; Weiss & Yoes, 1991). The point-biserial correlation can be defined as:

$$\rho_{pb} = \frac{\mu_1 - \mu_x}{\sigma_x}\sqrt{b/q}\,, \tag{22}$$

where μ_1 is the mean of the observed score among respondents who provide a correct response to the item, μ_x is the mean of the observed score for all respondents, σ_x is the standard deviation of the observed score for all respondents, *b* indicates the item difficulty index for the item, and *q* is 1-*b* (Suen, 1990). The ranges of point-biserial correlation are Very good (> .4), Good (< .39, > .3), Fair (< .29, > .2), Non-discriminating (< .19, > .0), and Needs attention (< .0). The optimal item discrimination is .5 (Alagumalai & Curtis, 2005). When an item exists in low item

discrimination, a correct response to the item has little or no relationship with the total score. Items with low or zero item discrimination will be removed from a test in order to improve the reliability of a test (Alagumalai & Curtis, 2005; Weiss & Yoes, 1991).

2.1.5 The Strength and Limitations of CTT

A weak theoretical assumption, which makes CTT easy to employ in many testing situations, is the important advantage of CTT. Other advantages of CTT include that smaller sample sizes are available for analyzing, its model parameter estimation is simple, and data analyses do not require stringent goodness of fit studies in order to ensure a good fit of the model to the measurement data. However, CTT includes several limitations. First, respondent- or test-characteristics are dependent (Hambleton et al., 1991). A respondent's observed score is dependent on the specific scale or test used, that is, test- or scale-dependent. When the test is difficult, the respondents will gain low true scores, whereas when the test is easy, they will have high true scores. Thus, when the test context is changed, respondents' predicted true scores change as well. Second, the person statistics are item dependent. Since the test scores in CTT are based on the total number-correct scores, respondents have scored dependence on the number of correct items. Thus, test scores depend on the item difficulties of the test selected (Weiss & Yoes, 1991). Third, the item statistics, that is item difficulties and item discriminations, are dependent on a sample where the test items were administered (Hambleton & Jones, 1993; Weiss & Yoes, 1991). For instance, the item parameters (item difficulties and item discriminations) would be different if the test items were administered to respondents from a high-ability group and from a low-ability group.

Although the total score of CTT is easy to compute and to understand, it is based on a weak assumption. Thus, it is difficult to obtain a consistency of difficulty, discrimination, and reliability on a similar test. Item response theory (IRT; Hambleton et al., 1991; Lord, 1952; Lord & Novick, 1968) overcomes these limitations.

2.2 The Basic Concept of Item Response Theory (IRT)

IRT is built on mathematical models and statistical methods that associate item responses with a latent trait and use it to analyze items and scales. This latent trait is a hypothetical variable that measures individuals on the psychological constructs of interest on a scale, such as depression scale. This scale is continuous, has equal intervals, and is free from measurement error. The latent trait is referred to as ability in the educational and psychometric fields and is denoted by theta (θ). Ability is assumed to underlie the observed responses for a set of items (Osgood, McMorris, & Potenza, 2002), and using either one or more item properties, such as item difficulty, describes an item. IRT is a statistical theory related to a respondent's ability and performance on a test, and abilities are measured by the test items (Hambleton & Jones, 1993). Thus, IRT mainly focuses on item-level information. It has been developed for tests whose items are scored dichotomously, such as yes (1) or no (0) and polytomously, such as short answer tests scored 0, 1, or 2 (Kolen & Brennan, 2004). The responses on items can be discrete or continuous, and the categories of item scores can be ordered or unordered.

2.2.1 Assumptions

The mathematical models of IRT specify that a respondent's probability of giving a correct item response is dependent upon the respondent's ability and an item's characteristic.

Because IRT is based on a family of the mathematical models, strong assumptions must be made under which the model can assume to be held (Weiss & Yoes, 1991). The first assumption is dimensionality of the latent spaces (or variables). Dimensionality includes unidimensionality and multidimensionality. However, most applications of IRT models have assumed unidimensionality of latent space. The unidimensional model, which requires that tests measure only a single ability, is defined as a model intended to measure a single dimension. Thus, it is appropriate for a single common factor for item response. Under the unidimensional IRT model for dichotomously scored tests, a respondent's ability is described by a single latent trait referred to as θ, and the range of θ is $-\infty < \theta < \infty$ (Kolen & Brennan, 2004). Nevertheless, when a test is administered under at least two dimensions, such as speeded tests measuring response speed and latent trait, unidimensional IRT models are not appropriate for more than one dimensional test. The development of the multidimensional IRT model is important. The multidimensional model is defined as a model intended to measure two or more dimensions. The multidimensional model is appropriate for two or more abilities to have a different impact on the items, and it is appropriate for respondents who have different systematic strategies, knowledge structures, or interpretations to apply to the items (Embretson & Reise, 2000). Nonetheless, the multidimensional IRT model is more complex and is not commonly applied.

The second assumption is local independence (LI). LI means that respondents' responses to any set items on a test are statistically independent after taking the ability of a respondent into account (Hambleton et al., 1991; Kolen & Brennan, 2004). This implies that the ability and response of a respondent to the items are independent. Given the ability of a respondent to answer the items, the probability under the LI is equal to the product of the probability of answering each individual item. The LI can be defined as:

$$P\left(Y_1, Y_2, ..., Y_n \mid \theta_j\right) = \prod_{i=1}^{n} P\left(Y_i \mid \theta_j\right), \quad (23)$$

where $P\left(Y_1, Y_2, ..., Y_n \mid \theta_j\right)$ is the probability of a response pattern on *n* items given the ability for respondent *j* (Hambleton et al., 1991). A concept of the LI is related to unidimensionality if the IRT model contains respondents' abilities only on a single dimension.

Third, the item characteristics curve (ICC) is the main feature of IRT analyses, and it has a specified form. The form of the ICC describes the relationship between the probabilities of a correct item response and respondent's ability level on the construct being measured by the item. This relationship is characterized by the item difficulty on the ability scale and by its discrimination (Baker & Kim, 2004; Suen, 1990). Without this relationship, items cannot be differentiated between respondents with high and with low ability levels. The ICC specifies that when the ability level of a respondent increases, the probability of a correct response on items increases. In addition, it can be defined by such a mathematical function as $P(\beta_i, \alpha_i, \gamma_i, \theta) \equiv P_i(\theta)$. θ is a latent trait, which is referred to as ability. $P_i(\theta)$ represents the probability of a correct response at any point on the θ scale, and *i* represents an item ($i = 1, 2, \ldots, n$). β_i is the difficulty parameter, α_i is the discrimination parameter, and γ_i is the lower asymptote (Baker & Kim, 2004). Note that item difficulty, item discrimination, and lower asymptote were referred to as Greek letters *β*, *α*, and *γ* or as *b*, *a*, and *c* parameters, respectively in this study. Figure 2 shows that when the probability of giving the correct response is .5, the ability on the scale is zero. When the probability is 1, the ability is 4, whereas when the ability is -4, the probability is zero as in Item 3.

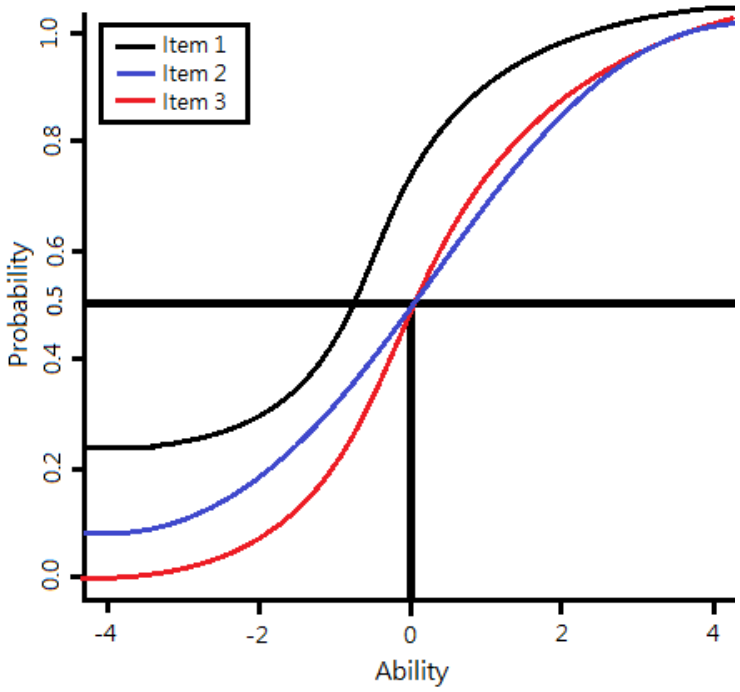

Figure 2. The item characteristic curve for three dichotomous items.

The ICC includes three features (Embretson & Reise, 2000). First, the ICC is S-shaped, which plots the probability of correct response for the respondent to the item as a monotonically increasing function of ability. Second, the shape displays that items differ in location, slope, and lower asymptote. Location relates to item difficulty in the ICC, and it describes the extent to which items are different in probabilities across ability. Slope, which refers to item discrimination, describes how rapidly the probabilities change with ability. The change of Item 1 and Item 3, for example, were much faster than Item 2; thus, Item 1 and Item 3 were more discriminating than Item 2, because probabilities of item response were relatively more responsive to changes in ability level. The lower asymptote means that the probability of success does not fall to zero no matter how low ability. For instance, the range of the probabilities of Item 2 and 3 were between .0 and .2 shown in Figure 2. However, the range of the probability of

Item 1 was over .2, and this means that the item could be solved by guessing with the lower asymptote larger than zero no matter how low ability. Three items had different lower asymptotes but they had the same difficulty level of 0, indicated in the vertical line shown in Figure 2. For example, the probability of a correct response of Item 1 was higher than .5 with lower asymptotes of .25. Third, if a respondent is considered to have a low ability level, the probability of a correct response to specific items will be close to zero. On the other hand, if a respondent is considered to have a high ability level, the probability of a correct response to specific items will be close to 1.0.

2.2.2 The Basic Components of IRT

In a sense, some concepts and some terminologies of IRT are similar to the CTT. For example, the concept of item parameters, such as item difficulty and item discrimination, is applied in both IRT and CTT. Even though the definitions of item parameters are different, the basic ideas are similar. However, the differences between IRT and CTT are that IRT includes three fundamental concepts: item response functions, information functions, and invariance (Reise, Ainsworth, & Haviland, 2005; Suen, 1990; Weiss & Yoes, 1991).

2.2.2.1 Item Response Functions (IRFs)

Item response functions (IRFs) are the basic unit of IRT. They describes the relation between a latent variable, which is individual differences on a construct such as depression, and the probability of correct item response to measure the latent variable. IRFs are normally used to evaluate item quality and serve as building blocks in order to get important psychometric properties. Three parameters, item difficulty, item discrimination, and lower asymptote, describe

the different aspects of the IRFs. For instance, Figure 3 shows the three dichotomously scored items (e.g., Yes or No) of the IRFs. When the location changes along with the ability axis, the IRF curve changes its inflection point given the difficulty of an item. Thus, Item 3 is more difficult than the other Items. The parameter of item difficulty is similar to the item mean in CTT. The steepness at the inflection point of the curve is called item discrimination. An item can be better to discriminate between a respondent having ability range around an item difficulty when having more discriminating items. Item 2 shown in Figure 3, for example, is more discriminating than the other two items. The parameter of item discrimination is similar to the item-test correlation in CTT. The lower asymptote that illustrates the IRFs is the probability related to the lower bound of the curve. In addition, it represents the probability of giving a correct response for individuals with low ability; thus, this parameter is referred to as the pseudo-chance level parameter (Reise et al., 2005; Weiss & Yoes, 1991). For instance, Item 1 has the

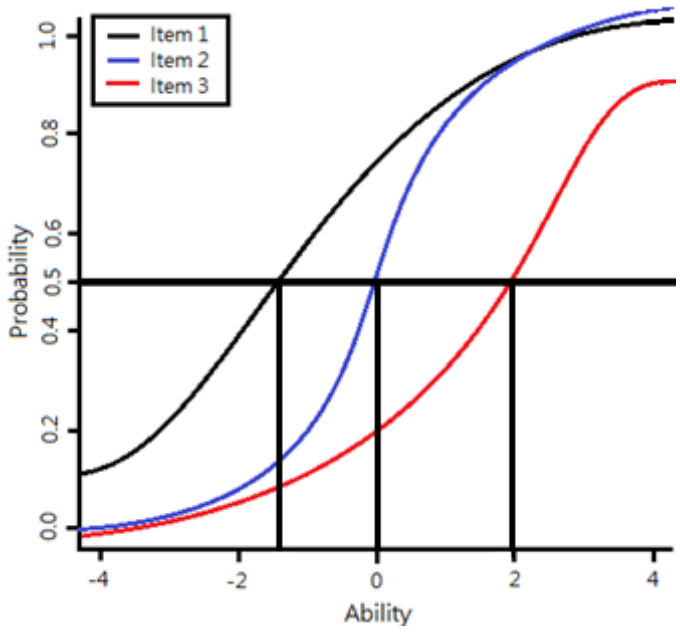

Figure 3. Item response functions for three dichotomously scored items.

lower asymptote in Figure 3.

When applying three parameters in test items, a three-parameter model has been used. Assuming that the lower asymptote is equal to zero, and applying only the item difficulty and item discrimination, a two-parameter model is being used. When applying only item difficulty in a test item, assuming all item discriminations are the same or setting them to 1 and the lower asymptote to zero, the one-parameter or Rasch model has been used. The normal ogive and logistic ogive models can be used to describe the IRFs.

2.2.2.1.1 Normal Ogive Model

The unidimensional two-parameter normal ogive model can be defined as

$$P_i(\theta) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \tag{24}$$

where $P_i(\theta)$ is the probability of the item response for a respondent, and z_i can be written as

$$\frac{\theta - \mu_i}{\sigma_i} = \frac{1}{\sigma_i}(\theta - \mu_i) = \alpha_i(\theta - \beta_i), \tag{25}$$

where θ is the ability, μ_i is the mean, and σ_i is the standard deviation, which is a measure of the spread of the normal distribution. When σ_i is large, the normal ogive is flat. On the other hand, when the middle section of a normal ogive is steep, the value of σ_i is small. $1/\sigma_i$ is equal to α_i that represents the discrimination parameter, and μ_i is equal to β_i, which is the difficulty parameter (Baker & Kim, 2004).

Assuming that all α_i are the same to become a constant value, $\alpha_i(\theta - \beta_i)$ from Equation 25 will become $\theta - \beta_i$. Since only one parameter (β_i) will be estimated, this is a one-parameter

normal ogive model (Suen, 1990) and equation can be simply written as

$$P_i(\theta) = \int_{-\infty}^{\theta-\beta_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz\,. \tag{26}$$

Equation 24 also can be extended to a three-parameter normal ogive model, and the equation can be written as:

$$P_i(\theta) = \gamma_i + (1-\gamma_i) \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \tag{27}$$

where γ_i parameter is the lower asymptote (Suen, 1990). Since the normal ogive model is more complex, Birnbaum (1968) is demonstrating a mathematically more convenient model, a logistic ogive model.

2.2.2.1.2 Logistic Ogive Model

Unlike the normal ogive model, $P_i(\theta)$ of the logistic model can be computed directly because the logistic model does not involve integration; thus, it is more popular to use in IRT. This model also includes three popular models that are the one-parameter logistic (1PL) model or Rasch model, two-parameter logistic (2PL) model, and three-parameter logistic (3PL) model.

1PL model is the simplest IRT model for binary response data. The discrimination parameter is fixed for all items, and only the difficulty parameter can compute on different values. It is equivalent to what is known as the Rasch model where the discrimination parameter is set to a value of 1.The equation for the 1PL model can be written as:

$$P(Y_{ij} = 1 \mid \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{28}$$

where $P\left(Y_{ij}=1|\theta_j,b_i\right)$ is the probability of correct item response for respondent *j* to item *i*; and b_i is the item difficulty parameter of item *i*, and it is the point on the ability scales where the probability of a correct item response from respondent *j* to item *i* is .5 (Baker & Kim, 2004; Hambleton et al., 1991).

By contrast, the 2PL model includes the item discrimination a_i and the 2PL can be written as:

$$P\left(Y_{ij}=1|\theta_j,b_i,a_i\right)=\frac{\exp\left[a_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[a_i\left(\theta_j-b_i\right)\right]}, \tag{29}$$

where a_i is the item discrimination parameter, and it describes how well an item can discriminate between a respondent having an ability level below or above the item location. a_i reflects the steepness of the ICC in its middle section. The steeper the curve, the more the item can be discriminated. On the other hand, the flatter the curve, the less the item can be discriminated, because the probability of a correct item response at low ability levels is nearly the same as at high ability levels (Baker, 2001).

Under the 3PL model, Birnbaum (1968) modified the 2PL to include a parameter, c_i, which is a lower asymptote. The 3PL can be written as:

$$P\left(Y_{ij}=1|\theta_j,b_i,a_i,c_i\right)=c_i+\left(1-c_i\right)\frac{\exp\left[a_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[a_i\left(\theta_j-b_i\right)\right]}, \tag{30}$$

where c_i is the lower asymptote, and it is the probability of respondents with low abilities correctly answering an item (Baker & Kim, 2004). The difficulty parameter is the point on the ability scale where $P\left(\theta\right)=\left(1+c\right)/2$, and the discrimination parameter is proportional to the slope

that is $a(1-c)/4$ of the ICC at $\theta = b$ (Baker, 2001).

2.2.2.2. Information Functions

The concept of information function is crucial in IRT. Information function is an index indicating the item's ability to distinguish among individuals. (Reise et al., 2005; Weiss & Yoes, 1991). Sir R. A. Fisher (1922) described that information function is defined as the reciprocal of the precision with which an item parameter is measured. The amount of information, *I*, is given by

$$I = \frac{1}{\sigma^2}, \tag{31}$$

where σ^2 is the variance of the estimators. The amount of information in a test can be computed for each ability level on the ability scale from $-\infty$ to ∞. The information function tells us how well each θ level is being estimated. When *I* is large, true ability can be estimated with precision. When *I* is small, ability cannot be estimated, and the estimates will be widely scattered about the true ability.

The IRFs can be transformed to the item information function (IIF), which is described as the relationship between an item's informativeness and ability as shown in Figure 4. The amount of information based on a single item can be computed at an ability level. We will get different amounts of information in a different range of a given ability from different items. The relatively easy items are useful for discriminating among individuals on low ability, whereas the relatively difficult items are useful for discriminating among individuals on high ability (Reise et al., 2005). Figure 4, for example, shows that Item 2 provides different item difficulty of the amount

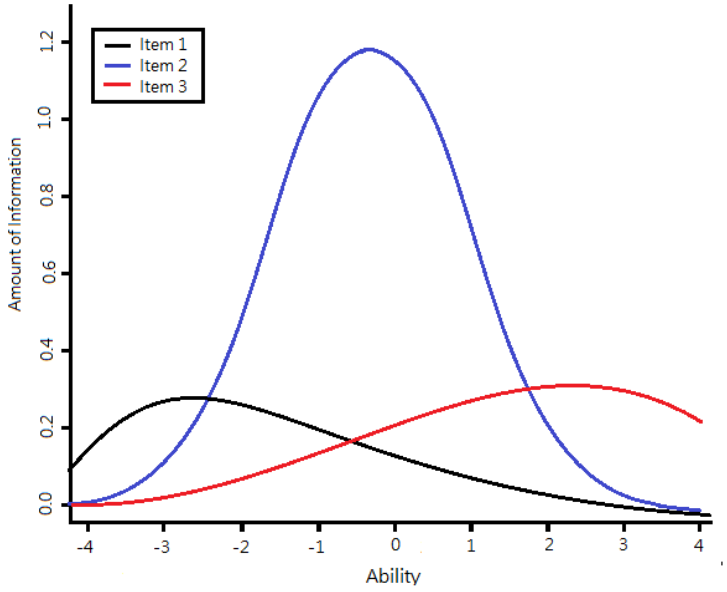

Figure 4. Item information function for three example items.

of information in a different ability range, and it provides more information related to discriminating items.

The IIF at a specific ability value is conceptually a ratio of the slope of the ICC and the expected measurement error at the specific ability, and the equation is denoted as

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}, \tag{32}$$

where $I_i(\theta)$ is the amount of information for item *i* at the θ level; $P_i' = \frac{\partial P_i(\theta)}{\partial \theta}$, that is the first derivative of the ICC. $P_i(\theta)$ is the probability of a correct response, and it depends on the particular ICC model used; $Q_i(\theta) = 1 - P_i(\theta)$, that is the probability of the incorrect response. $P_i'(\theta)$ can be substituted in the actual derivatives for the three popular logistic models (Baker & Kim, 2004).

The IIF under the 1PL or Rasch model can be written as

$$I_i(\theta) = P_i(\theta)Q_i(\theta), \tag{33}$$

where $P_i(\theta) = \frac{1}{1+\exp^{-(\theta - b_i)}}$, $Q_i(\theta) = 1 - P_i(\theta)$.

The IIF under the 2PL model can be defined as

$$I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta), \tag{34}$$

where a_i is discrimination parameter, $P_i(\theta) = \frac{1}{1+\exp^{-a_i(\theta - b_i)}}$.

The IIF under the 3PL model can be written as

$$I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta)\left[\frac{P_i^*(\theta)}{P_i(\theta)}\right]^2, \tag{35}$$

where $P_i^*(\theta) = \frac{1}{1+\exp^{-a_i(\theta - b_i)}}$, $P_i(\theta) = c_i + (1 - c_i)\frac{1}{1+e^{-a_i(\theta - b_i)}}$.

A test is formed by a series of items, and each item has its own IIF. When the IIF from different items adds together with the assumption of local independence, this forms a test information function (TIF) denoted by $I(\theta)$ (Baker & Kim, 2004; Birnbaum, 1968; Suen, 1990). Figure 5 shows the TIF for three example items from Figure 4. The TIF tells us how well the test is doing in estimating θ over the whole range of θ scores. The equation of TIF is defined as

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \tag{36}$$

where $I(\theta)$ is the amount of test information at the ability level; $I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}$; and n is the number of items in the test.

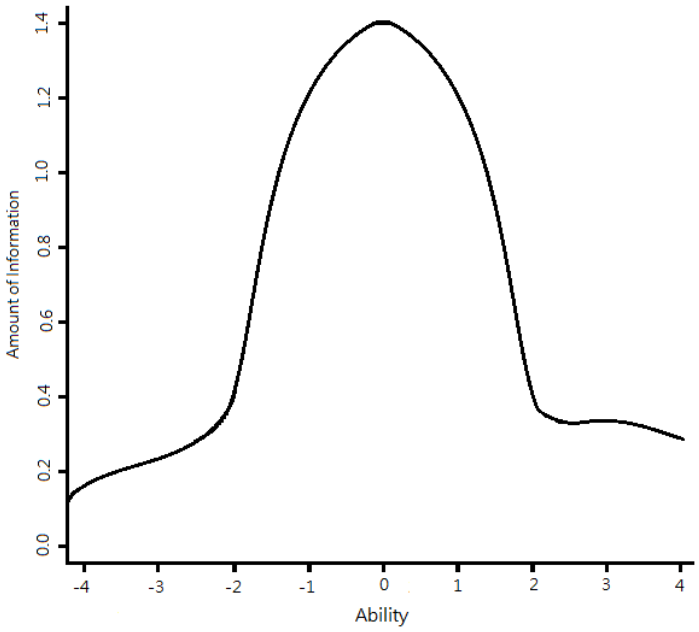

Figure 5. Test information function for three example items.

The item and test information are similar to the item and test reliability of CTT. Nevertheless, the major difference is that information in IRT can differ depending on respondents who fall along a certain ability range, while the scale reliability in the CTT is the same for all respondents, no matter what their raw-score levels (Reise et al., 2005).

2.2.2.3. Invariance

The values of item parameters in IRT model are constant and do not depend on the characteristics of a sample population when item parameters are measured. In other words, the item parameters in IRT model have an invariance property within a linear transformation, and this is known as item-parameter invariance. Unlike the CTT's *p*-values, which change when the different sample populations are used, item parameters in IRT will maintain the same values, no matter what sample population has been used. In addition to item-parameter invariance, the

ability scale does not depend on a specific set of items. However, in CTT, the raw score scale is dependent on a specific item set on a single measure. The advantages of item-parameter invariance allow researchers to efficiently link diverse scales that estimate the same latent variable and to compare respondents who have responded to different items (Reise et al., 2005; Suen, 1990).

2.2.3 The Method of IRT Test Scoring

The major goal of a test is to obtain a measure of the ability of each respondent. The estimated ability of the respondent in IRT is not the same as in CTT. In CTT, total test score is obtained by summing correct responses to the items, and this summed score will linearly transform to a scale score in order to estimate ability level of a respondent. In contrast, the IRT score is a non-linear function of the manifest item responses. It uses the item parameters and the knowledge of how these item parameters affect the ICC to estimate the respondent's ability score based on his or her item response. The estimated ability for the respondent has maximized the likelihood of the respondents' item response patterns given the scores of the parameter of *n* dichotomously scored test items (Baker & Kim, 2004; Embretson & Reise, 2000). In a binary test item, if a respondent provides a correct answer, his or her response is scored as a 1, otherwise scored 0. Since ability is randomly put on the standard *z*-scale, the score of ability may range roughly from -4 to 4. IRT is used to convert item response into a scale to estimate ability and to calibrate items and measure item parameters.

The probability of giving a correct item response is denoted as $P(\theta)$, and the probability of giving an incorrect item response is denoted as $Q(\theta)$ [i.e.,$1-P(\theta)$]. The probability of giving correct or incorrect responses yields a function that is monotonically

increasing or decreasing. In addition, the probability of correct item response in a test is independent of the probability of correct or incorrect responses to the remaining items under the assumption of local independence. The function described by multiplying the probabilities of correct and incorrect response patterns is known as the likelihood function (LF), which is crucial to all IRT parameter estimation (Embretson & Reise, 2000; Weiss & Yoes, 1991).

For the LF, suppose that a randomly given respondent *j* responds to the *n* items with response pattern defined $U_j = (u_{1j},\ldots,u_{nj})$ to be observed responses where *j* represents the respondent $j = 1,\ldots,N$, and u_{ij} is dichotomously scored item (0 or 1), $i = 1,\ldots,n$. For instance, we multiply the three item IRFs in order to get the maximum likelihood for the item response pattern. If a respondent answers 1, 1, 0 for three examples items shown in Figure 6, the IRFs, $P_1(\theta)$, $P_2(\theta)$, and $Q_3(\theta)$, will be multiplied together and the conditional likelihood will be obtained. The LF of the probability of the vector of observed responses for a given respondent *j*

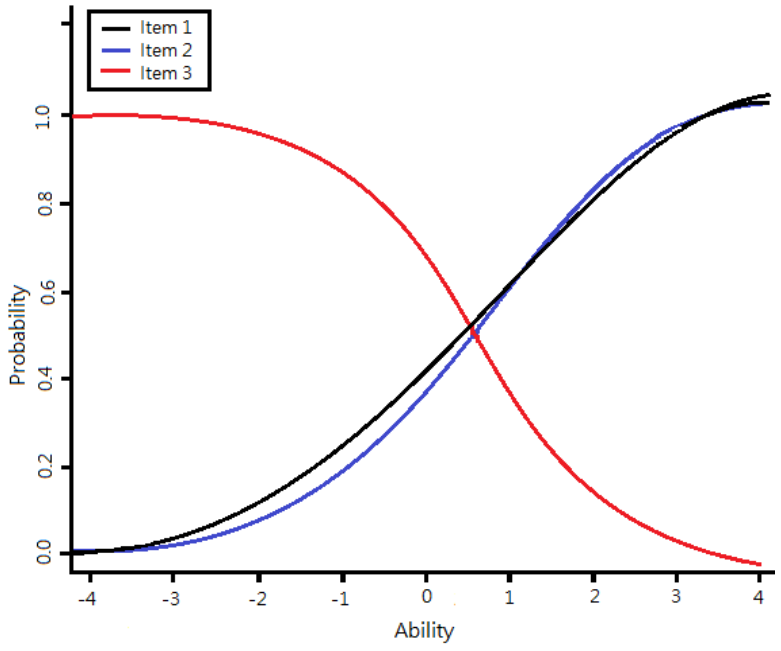

Figure 6. Item response functions for three dichotomous item.

can be computed by

$$\ell(\theta_j) = \text{Prob}\left(U_{jn} \mid \theta_j\right) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}}, \quad (37)$$

where $Q_i\left(\theta_j\right) = 1 - P_i\left(\theta_j\right)$ (Baker & Kim, 2004; Embretson & Reise, 2000). The likelihood is low for a low score of θ because it is unlikely that a respondent answers Item 1 correctly, whereas the likelihood is low for the high score of θ because it is unlikely that a respondent answers Item 3 incorrectly. Thus, the likelihood for these three items is shown in Figure 7, that about $\theta = 0.6$ is the best. When one obtains the highest likelihood, maximum likelihood estimation can be used to predict the score of θ for a respondent.

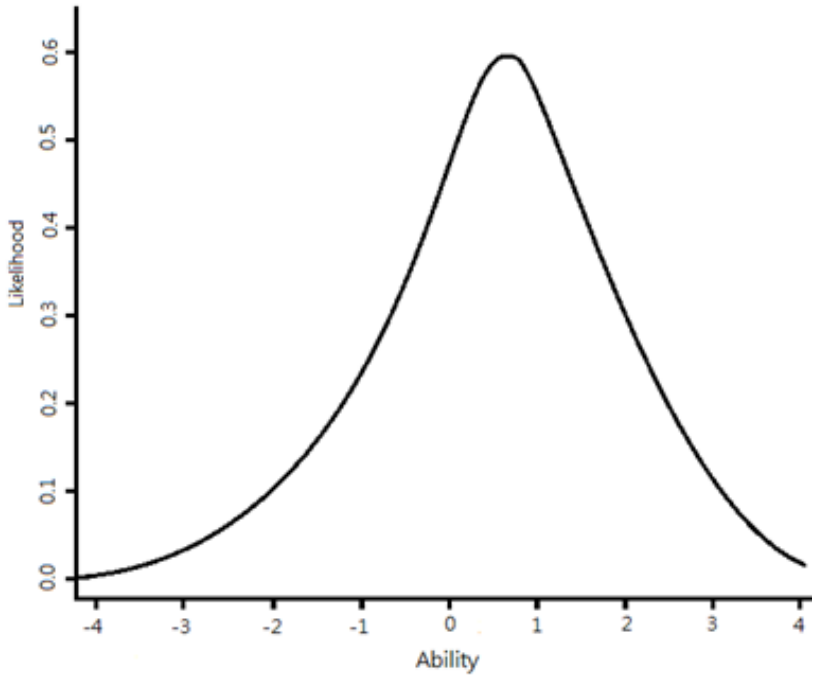

Figure 7. The likelihood function for three dichotomous items.

2.2.3.1 Maximum Likelihood Estimation

In IRT, ability is estimated in a model (i.e., 1PL, 2PL, or 3PL) for the response of a respondent after controlling for the item properties. The ability can be estimated by a maximum likelihood method. The maximum likelihood estimation (MLE) is used to estimate the ability of a respondent based on searching the ability score that maximizes the LF of a participant's response pattern given item parameters. There are three assumptions to achieve the MLE (Baker & Kim, 2004). First, the scores of the item parameters are known. Second, the respondents must be independent, and their abilities can be estimated on a respondent by respondent basis. Third, all items on the test are modeled by the ICC models, that is 1PL, 2PL, or 3PL.

The maximum likelihood procedure is an iterative process, and it begins with some a priori value for a respondent's ability and known item parameter (Baker, 2001). The probability of obtaining a response u_{ij}, which is the observed item response of a respondent *j* to item *i*, for a dichotomous score to the *n* items, given the respondent θ_j and the item parameters $\xi_i = (\lambda_i, \zeta_i)$, can be written as

$$P(u_{ij} \mid \theta_j, \xi_i) = P_i(\theta_j)^{u_{ij}} Q_i\left(\theta_j\right)^{1-u_{ij}}, \tag{38}$$

where $P_i(\theta_j) = \dfrac{1}{1+\exp[-(\zeta_i + \lambda_i\theta_j)]}$, and assumes λ_i and ζ_i parameter are known. ζ_i is an intercept, and λ_i is a slope (i.e., the discrimination parameter); $Q_i\left(\theta_j\right) = 1 - P_i(\theta_j)$. Thus, the slope/intercept form, $\zeta_i + \lambda_i\theta_j$, will be used in the estimation procedures (Baker & Kim, 2004).

The MLE treats the likelihood as a function of θ_j and attempts to obtain the score of θ_j that maximizes the likelihood. However, one of the problems of the MLE is the IRFs include the scores between zero and 1 that are multiplied together in Equation 38. Hence, the conditional

likelihood can become very small, and a computer program loses precision. In order to solve this problem, we may work with log-likelihood instead of the raw likelihood (Baker & Kim, 2004; Embretson & Reise, 2000).The log-likelihood is calculated by summing the IRFs instead of multiplying the IRFs, using the natural logarithm of the IRTs. The log-likelihood can be computed by,

$$L = \log \ell(\theta_j) = \sum_{i=1}^{n} [u_{ij} \log P_i(\theta_j) + (1-u_{ij}) \log Q_i(\theta_j)], \tag{39}$$

where u_{ij} is the observed responses. Because all item parameters are known, the first and the second derivatives of the log-likelihood with respect to a given respondent can be computed. Note that the derivatives of $P_i(\theta_j)$ and $Q_i(\theta_j)$ with respect to θ_j will be dependent on the particular ICC model employed. For instance, letting $P_{ij} = P_i\left(\theta_j\right)$ and $Q_{ij} = Q_i\left(\theta_j\right)$, the first derivative for 2PL of the log-likelihood function with respect to θ_j can be computed with Equation 39,

$$\begin{aligned}
\frac{\partial L}{\partial \theta_j} &= \frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^{n} [u_{ij} \log P_i(\theta_j) + (1-u_{ij}) \log Q_i(\theta_j)] \right\}, \\
&= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} [u_{ij} \log P_i(\theta_j)] + \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \left[(1-u_{ij}) \log Q_i(\theta_j) \right], \\
&= \sum_{i=1}^{n} u_{ij} \frac{1}{P_{ij}} \frac{\partial P_i\left(\theta_j\right)}{\partial \theta_j} + \sum_{i=1}^{n} \left(1-u_{ij}\right) \frac{1}{Q_{ij}} \frac{\partial Q_i\left(\theta_j\right)}{\partial \theta_j}.
\end{aligned} \tag{40}$$

Applying the chain rules to calulate $\frac{\partial P_i\left(\theta_j\right)}{\partial \theta_j}$, the Equation of $\frac{\partial P_i\left(\theta_j\right)}{\partial \theta_j}$ can be written as

$\frac{\partial P_i\left(\theta_j\right)}{\partial \theta_j} = f'g\left(\theta_j\right)g'\left(\theta_j\right)$, letting $f\left(\theta_j\right) = \frac{1}{g\left(\theta_j\right)}$ and $g\left(\theta_j\right) = 1 + \exp\left[-\alpha_i\left(\theta_j - \beta_i\right)\right]$.

$$g'(\theta_j) = \frac{\partial}{\partial(\theta_j)}\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)$$

$$= \frac{\partial}{\partial(\theta_j)}\left[1+\exp\left(-\alpha_i\theta_j+\alpha_i\beta_i\right)\right]$$

$$= -\alpha_i \exp\left[-\alpha_i(\theta_j-\beta_i)\right]. \tag{41}$$

$$f'g(\theta_j) = \frac{\partial}{\partial(\theta_j)}\frac{1}{g(\theta_j)} = \frac{\partial}{\partial(\theta_j)} g(\theta_j)^{-1}$$

$$= -g(\theta_j)^{-2}$$

$$= -\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)^{-2}. \tag{42}$$

When combining Equations 41 and 42, $\frac{\partial P_i(\theta_j)}{\partial \theta_j}$ can be computed by

$$\frac{\partial P_i(\theta_j)}{\partial \theta_j} = -\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)^{-2} \times \left(-\alpha_i \exp\left[-\alpha_i(\theta_j-\beta_i)\right]\right)$$

$$= \frac{\alpha_i \exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]}{\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)^2}$$

$$= \alpha_i \times \frac{1}{\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)} \times \frac{\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]}{1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]}$$

$$= \alpha_i \times \frac{1}{\left(1+\exp\left[-\alpha_i\left(\theta_j-\beta_i\right)\right]\right)} \times \frac{1}{1+\exp\left[\alpha_i\left(\theta_j-\beta_i\right)\right]}$$

$$= \alpha_i P_i(\theta_j) Q_i(\theta_j). \tag{43}$$

$$\frac{\partial Q_i(\theta_j)}{\partial \theta_j} = -\alpha_i P_i(\theta_j) Q_i(\theta_j). \tag{44}$$

Substituting Equation 43 and 44 into Equation 40 yields

$$= \sum_{i=1}^{n} u_{ij} \frac{1}{P_{ij}} \left[\alpha_i P_i(\theta_j) Q_i(\theta_j) \right] + \sum_{i=1}^{n} (1 - u_{ij}) \frac{1}{Q_{ij}} \left[-\alpha_i P_i(\theta_j) Q_i(\theta_j) \right],$$

$$= \sum_{i=1}^{n} \alpha_i \left[u_{ij} Q_i(\theta_j) + (1 - u_{ij}) \left(-P_i(\theta_j) \right) \right],$$

$$= \sum_{i=1}^{n} \alpha_i \left(u_{ij} Q_i(\theta_j) - P_i(\theta_j) + u_{ij} P_i(\theta_j) \right),$$

$$= \sum_{i=1}^{n} \alpha_i \left[\left(u_{ij} Q_i(\theta_j) + u_{ij} P_i(\theta_j) \right) - P_i(\theta_j) \right],$$

$$= \sum_{i=1}^{n} \alpha_i \left[u_{ij} \left(1 - P_i(\theta_j) + P_i(\theta_j) \right) - P_i(\theta_j) \right],$$

$$= \sum_{i=1}^{n} \left[\alpha_i \left(u_{ij} - P_i(\theta_j) \right) \right]. \quad (45)$$

The second derivative for 2PL can be computed with Equation 45,

$$\frac{\partial^2 L}{\partial \theta_j^2} = \frac{\partial}{\partial \theta_j} \left\{ \sum_{i=1}^{n} \left[\alpha_i \left(u_{ij} - P_i(\theta_j) \right) \right] \right\}.$$

$$= \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \alpha_i u_{ij} - \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \alpha_i P_i(\theta_j),$$

$$= -\sum_{i=1}^{n} \alpha_i \frac{\partial}{\partial \theta_j} P_i(\theta_j),$$

$$= -\sum_{i=1}^{n} \alpha_i \left[\alpha_i P_i(\theta_j) Q_i(\theta_j) \right],$$

$$= -\sum_{i=1}^{n} \alpha_i^2 P_i(\theta_j) Q_i(\theta_j). \quad (46)$$

Equation 39 is useful in examining a short test. However, since some data sets may contain thousands of respondents answering many items, such as 50 or more items, one needs a

numerical method to find the exact maximum log-likelihood given a particular item response pattern. A Newton-Raphson procedure is the most commonly used method to compute the maximum of the log-likelihood. It will be employed to obtain the estimate of an ability parameter through iteration. The Newton-Raphson algorithm is used to find the mode of log-likelihood of each respondent. The first procedure in the Newton-Raphson scoring algorithm is specified ability starting value. When obtaining the tentatively estimated ability value, we can calculate the first (i.e., Equation 40) and second (i.e., Equation 46) derivatives of the log-likelihood function using this ability value. By obtaining these ability values, we can compute the ratio, that is, the first derivative divided by the second derivative. Thus, the Newton-Raphson iteration can be defined as (Baker & Kim, 2004)

$$\left[\hat{\theta}_j\right]_{t+1} = \left[\hat{\theta}_j\right]_t - \left[\frac{\partial^2 L}{\partial \theta_j^2}\right]_t^{-1} \left[\frac{\partial L}{\partial \theta_j}\right]_t, \tag{47}$$

where t indexes the iteration, and $\left[\hat{\theta}_j\right]_t$ is the estimated ability of the respondent within iteration t. The iteration will continue until a stable set of parameter estimates is obtained. Using this Newton-Raphson iteration, we can find a new value of ability.

The MLE has several features. First, since the expected value of ability is always equal to the true ability, it is unbiased. Second, the MLE is a proficient estimator, and its errors are based on normal distribution. Nevertheless, the main problem of the MLE is that when a respondent answers all items correctly (i.e., perfect response vectors) or incorrectly (i.e., zero response vectors), the MLE process cannot provide estimates of ability for him or her, because the LR is unable to have a single identifiable peak. In addition, MLE works best only with large sample sizes and large item pools, such as 50 items (Embretson & Reise, 2000). Moreover, the MLE may be incomputable under certain conditions in 3PL. In this situation, Bayesian model

estimation is useful to solve the limitation of the maximum likelihood method.

2.2.3.2 Bayesian Model Estimation

Bayesian model estimation (BME) assumes that if we know the distribution of ability, we can use this information in making more accurate ability estimation. The assumed distribution of ability is called the prior distribution. When the prior distributions are used to estimate the ability level of a respondent, this process is also known as the maximum a posteriori (MAP) scoring strategy (Embretson & Reise, 2000).

MAP estimation includes several concepts. The first concept is the notion of prior distribution. It is a hypothetical probability distribution; a researcher assumes respondents who are a random sample. Thus, the prior distribution in ability level estimation is the standard normal distribution, that is, respondents are sampled from a normal distribution with mean, μ = 0, and variance, σ^2 = 1. The second concept is log-likelihood. The third concept is the posterior distribution. The definition of posterior distribution is the LF, multiplied at each ability level by the density of the prior distribution at that same ability level. The equation of posterior distribution can be written as (Hambleton & Jones, 1993; Suen, 1990)

$$f(\theta|U) \propto L(U|\theta)g(\theta), \tag{48}$$

where $f(\theta|U)$ is the posterior distribution; $g(\theta)$ is the prior distribution; and $L(U|\theta)$ is the LR as in Equation 37.

The purpose of MAP is to obtain the score of ability that maximizes the posterior distribution that will equal the mode. To estimate respondent ability level using MAP scoring, which is specified as normal distribution, the procedure can be used the same as the MLE. That is, using a specific response pattern and a parameter of items in a test, we can compute the log-

likelihood and the first and second derivatives of LF (Embretson & Reise, 2000).

The strengths of the MAP are that it incorporates prior information, all perfect and zero responses vector can be scored, and the estimation of ability can be determined. However, it still includes some limitations (Embretson & Reise, 2000; Weiss & Yoes, 1991). The first limitation of the MAP is that scores may be biased in a short test, such as those with fewer than 20 items because the expected values of the estimation of MAP are not the same as the value of the true parameter. The second limitation is to assume a specific form for the prior distribution. If an incorrect prior has been used, scores are critically biased and misleading. The third limitation is that if the assumption of ability distribution is invalid, MAP will provide a poor result of ability estimation.

Besides MAP, we can also employ the expected a posteriori (EAP) to estimate ability level, and it is non-iterative. The EAP has provided a finite ability level estimate for all perfect and zero responses patterns. Unlike MAP, which is finding the mode of the posterior distribution, EAP is found by the mean of the posterior distribution. The scoring strategy of the EAP is that a set of probability densities on each set of test items is computed at a finite number of specified values of ability that is called quadrature nodes. These densities are taken from a normal distribution, and the equation of the normal distribution can be written as (Embretson & Reise, 2000):

$$F(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right) e^{\left(-\theta^2/2\right)}. \tag{49}$$

Each quadrature node (X_r) of the densities is called weights [$W(X_r)$], which serves as discrete prior distribution, and the weights will transform so that their sum equals 1. While establishing the quadrature nodes and weights, the EAP estimate of ability can be defined as

(Bock & Mislevy,1981)

$$E\left(\theta_j \mid U_j, \zeta\right) = \bar{\theta}_j = \frac{\sum_{r=1}^{q} X_r L\left(X_r\right) W\left(X_r\right)}{\sum_{r=1}^{q} L\left(X_r\right) W\left(X_r\right)}, \tag{50}$$

where $E\left(\theta_j \mid U_j, \zeta\right)$ is the unconditional expectation of θ_j given response vector U_j and item parameter ζ; X_r indicates the value of ability of the *q* quadrature nodes; $L\left(X_r\right)$ is the exponent of the log-likelihood function to evaluate at each of the *q* quadrature nodes; and $W\left(X_r\right)$ is the weight of each quadrature.

The advantages of the EAP estimator include that its ability estimate is easy to obtain because EAP, which uses discrete prior not a continuous prior, is a non-iterative process. Thus, the EAP does not require one to know the first and second derivative of the LF. The EAP is simple to calculate, finds mean, and estimates finite ability level for all response patterns. However, the disadvantage of EAP estimator is that it is biased when existing in a finite number of items. The biased type is that the estimation of ability level is regressed toward the mean only if the number of items is large.

2.2.4 The Strengths and Limitations of IRT

The main advantages of IRT are its key features, such as ICCs, and strong assumptions, such as local independence. Besides, IRT includes other strengths (Hambleton & Jones, 1993). First, item statistics are independent on respondent samples from which they are obtained. Second, ability scores of respondents are independent on test difficulty. Third, test analysis does not need strict parallel tests to evaluate reliability. Fourth, test analysis provides matching test items to respondent knowledge level. Fifth, IRT provides detection of item bias and evaluation of

test score validity using differential item functioning (DIF) technique. IRT includes technical and practical limitations. On the technical limitations, IRT model is more complex than the CTT model. In addition, model fit is a problem because how model fit problems, such as a problem related to dimensionality on the test, should be addressed is not completely defined. On practical limitations, IRT often requires large sample in order to obtain precise and steady parameter estimates.

2.2.5 A Comparison between CTT and IRT (Embretson & Reise, 2000; Hambleton & Jones, 1993)

Classical Test Theory	Item Response Theory
CTT's model is a linear function of true score and random error.	IRT's model is a non-linear function of item parameters and ability.
CTT is focused on test-level information.	IRT is focused on item-level information.
CTT is referred to as a weak model because of the weak assumption that is easy to meet with measurement data.	IRT is referred to as a strong model because of strong assumptions, including dimensionality, local independence, and ICC; thus, IRT is less likely to meet with measurement data.
The true score has to mean only for a specific set of item properties because CTT does not include item properties in a model.	The ability has to mean for any set of calibrated items because IRT model includes item properties.
Item properties are not explicitly linked to behavior in CTT.	The relative impact of difficult items on ability estimates and item responses is known.
Ability and item parameters are the dependent variables.	Ability and item properties are independent variables if the model fits the data.
CTT only includes item difficulty and item discrimination.	IRT includes three parameters, item difficulty item discrimination, lower asymptote, and related to item information functions.
Item and respondent characteristics do not place on the common scale.	Item and respondent characteristics are placed on the common scale.
CTT is a nominal level measurement.	IRT is an interval level measurement, proving for Rasch model.
CTT has limited ability to detect measurement invariance.	IRT employs differential item functioning (DIF) to detect measurement invariance.
The required sample sizes for item parameter estimation are normally between 200 and 500.	Depending on the IRT models, IRT requires large sample sizes, normally over 500.

IRT model is widely used in order to measure latent ability, and it can also be used to analyze ordinal questionnaire data and to devise measurement scales. However, when researchers are studying growth or change using the IRT model, the traditional IRT model has to be extended with longitudinal data. The next section will discuss longitudinal item response theory.

2.3 Longitudinal Item Response Theory (LIRT) models

In constructing longitudinal research, it is necessary that the same construct is measured at all time points. This is in order that it is possible to observe change on the same construct (McArdle, Petway, & Hishinuma, 2014). However, researchers may ask: (1) How can they describe growth and change? How can they say that the structure is the same, but the score has changed over time? How can they attribute changes to the individuals but not changes to the scale of measurement? The longitudinal item response theory (LIRT) model is a way to answer these types of questions. LIRT employs an IRT model and allows researchers to investigate item and ability distribution parameters using longitudinal data (Tavares & Andrade, 2006). The LIRT model has received considerable attention for understanding changes in cognition or behavior over time (Choi, Harring, & Hancock, 2009; McArdle et al., 2014; Tavares & Andrade, 2006; von Davier, Xu, & Carstensen, 2011). Change across time points can be measured by focusing either on group differences or individual differences. In addition, it allows researchers to track students' progress over time in response to a new instructional treatment (Cho et al., 2013). The LIRT involves the measuring of group growth, the measuring of individual growth, and latent change score.

2.3.1 Measuring Group Growth

LIRT can be traced back to Fischer's (1973) linear logistic latent trait model (LLTM). LLTM is integrated with the linear regression model into the dichotomous Rasch model (RM). LLTM can be written as

$$P\left(Y_{ij}=1|\,\theta_j,\beta_i=\sum_{m=1}^{p}q_{im}\eta_m+c\right)=\frac{\exp\left(\theta_j-\beta_i\right)}{1+\exp\left(\theta_j-\beta_i\right)}$$

$$=\frac{\exp(\theta_j-\sum_{m=1}^{p}q_{im}\eta_m-c)}{1+\exp(\theta_j-\sum_{m=1}^{p}q_{im}\eta_m-c)}, \qquad (51)$$

where Y_{ij} represents binary response, $Y_{ij}=1$ when respondent *j* provides a correct response on item *i*; $P\left(Y_{ij}=1|\,\theta_j,\beta_i\right)$ is the probability of a correct response of respondent *j* on item *i*, given a respondent ability θ_j and item difficulty β_i. η_m, $m = 1, \ldots, p$, is the basic parameter, q_m is the given weight of the basic parameter *m* on item *i*, and *c*, which is a constant, represents the normalization (Fischer, 1997). The LLTM is a unidimensional model because it is a constrained RM. LLTM assumed that the treatment effects for all individuals measured were the same at the same time intervals; thus, it was not useful for measuring individual change (Wang, Kohli, & Henn, 2016). Fischer (1983) extended the LLTM model to the longitudinal case, namely a linear logistic latent trait model with relaxed assumptions (LLRA), for two-time points. The probability of LLRA at the first time point is dependent only on abilities and the second time point adds trend and treatment effects to the model (Embreston, 1991). The LLRA for two time points is

$$P\left(Y_{ij}=1|\,\theta_{ij},T_1\right)=\frac{\exp\left(\theta_{ij}\right)}{1+\exp\left(\theta_{ij}\right)}, \qquad (52)$$

$$P\left(Y_{ij}=1\,|\,\theta_{ij},T_2\right)=\frac{\exp\left(\theta_{ij}+\delta_j\right)}{1+\exp\left(\theta_{ij}+\delta_j\right)}, \tag{53}$$

and δ_j is decomposed into the sum of the basic parameter as follows:

$$\sum_{m=1}^{p} q_{jm}\eta_m+\tau\,, \tag{54}$$

where T_1 and T_2 are the first and the second time points, respectively; $P\left(Y_{ij}=1\,|\,\theta_{ij},T_1\right)$ and $P\left(Y_{ij}=1\,|\,\theta_{ij},T_2\right)$ are the probability of correct response for respondent *j* on item *i* given the item-specific ability parameter and the treatment effect at time points T_1 and T_2 , respectively; θ_{ij} is the item-specific ability of a respondent *j*; η_m is basic parameters that represent the treatment effect within T_1 and T_2; q_{jm} is the dose of treatment *m* given to θ_{ij} between T_1 and T_2; τ is the trend effect between T_1 and T_2 (Fischer, 1989, 1997). Although the LLRA model allows estimating structural parameters for treatments, it is not useful for measuring individual differences in responsiveness to treatment occasions (Embretson, 1991). In this regard, Andersen (1985) presented the multidimensional Rasch model for measuring individual differences to the repeated administration.

2.3.2 Measuring Individual Growth

Andersen's model assumed the exact same set of items to be administered across time points. Andersen (1985) developed this model to assess the effect of time on individual ability distribution. In other words, the ability (θ_j) of respondent *j* is extended to include time *t*. b_i is the item difficulty parameter of item *i*, and a latent linear score with time effects can be expressed as

$$P\left(Y_{ijt}=1\,|\,\theta_{jt}^{*},b_i\right)=\frac{\exp\left(\theta_{jt}^{*}-b_i\right)}{1+\exp\left(\theta_{jt}^{*}-b_i\right)}, \tag{55}$$

where $P\left(Y_{ijt}=1\,|\,\theta_{jt}^{*},b_i\right)$ is the probability of a correct response from respondent *j* on item *i* at the time point *t*, given the respondent ability θ_{jt}^{*} and item difficulty b_i. Y_{ijt} denotes the response of respondent *j* on item *i* at time point *t*, θ_{jt}^{*} is the ability of respondent *j* at time point *t*, and b_i denotes a time-invariant item difficulty parameter. In this model, item difficulties are constant across time points; however, the respondent's ability involves dependence on time points. In other words, different abilities, which might be independent, characterized time points. Although Andersen's model has taken into account time effects of the ability parameter, he did not explicitly consider change parameters for individuals. The ability in this model is time-specific, and it is unable to reflect individual differences in change over time points. Therefore, Embretson (1991) introduced a model for learning and change to examine individual differences in change over time.

The Embretson (1991) model, a multidimensional Rasch model for learning and change (MRMLC), estimates both initial ability and one or more modifiability (i.e., latent variables) from longitudinal data. This models the ability (θ_{jt}) of respondent *j* within occasions *t*. An initial ability (θ_{j1}) is involved in the item responses of respondent *j* at occasion 1, whereas a later ability (θ_{jt}) is involved in the item responses of respondent *j* at occasion *t* (*t* > 1). The MRMLC model can be defined as (Embretson, 1991)

$$P\left[Y_{ijt}=1\,|\,(\theta_{j1},\theta_{j2},...,\theta_{jt}),b_i\right]=\frac{1}{1+\exp\left(-\sum_{m=1}^{t}\theta_{jm}+b_i\right)}, \tag{56}$$

where θ_{j1} represents the initial level (t =1) ability of respondent j, $\theta_{j2}...\theta_{jt}$ are the tth modifiability, and it indicates the change in ability for respondent j between occasion t-1 to occasion t, and $\sum_{m=1}^{t}\theta_{jm}$ is the sum of an individual's initial ability $\left(\theta_{j1}\right)$and changes in the ability parameter $\left(\theta_{j(t-1)}\right)$. Item difficulty, b_i, is assumed to be the same for all occasions. Equation 56 shows that the MRMLC is a multidimensional Rasch model. In addition to the multidimensional model, the probability of item response can be given by a unidimensional model. The probability of giving correct item response is dependent upon the same composite ability θ_{jt}^{c} for all items within occasion t and for all t. θ_{jt}^{c} is the unweighted sum of the first ability and the t-1 modifiabilities up to time point t. Thus, for the unidimensional Rasch model, the probability of a correct response from respondent j on item i administered under occasion t is given as (Embretson, 1991)

$$P\left(Y_{ijt}=1|\,\theta_{jt}^{c},b_i\right)=\frac{1}{1+\exp\left(-\theta_{jt}^{c}+b_i\right)}, \tag{57}$$

where θ_{jt}^{c} indicates the ability of respondent j on occasion t and b_i is the difficulty of item i.

The difference between Andersen's (1985) model and Embretson's (1991) model is that the set of same items are repeated over time points in Andersen's model, whereas different sets of items are administered across time points in Embretson's model. If the same items are presented in the same test, this may lead to practice or/and memory effects to the same test taker. Thus, this may cause local dependency among item responses (von Davier et al., 2011).

Embretson (1997) further extended the MRMLC model to the 2PL model, namely structured latent trait models (SLTM). The SLTM is given as:

$$P\left(Y_{ijt}=1|\,\theta_{jm},b_{it},\lambda_{imt}\right)=\frac{1}{1+\exp\left(-\sum_{m=1}\lambda_{imt}\theta_{jm}+\sum_{t=1}b_{it}\right)}, \tag{58}$$

where $P\left(Y_{ijt}=1|\,\theta_{jm},b_{it},\lambda_{imt}\right)$ is the probability of the correct response of respondent *j* to item *i* under time point *t* given the ability of respondent θ_{jm}, item difficulty b_{it}, and the weight of ability λ_{imt}; θ_{jm} is the ability of respondent *j* on the weight of ability *m* where θ_{jm} is collected into a vector θ_j; b_{it} is item difficulty *i* under time point *t*, and λ_{imt} is the weight of ability *m* on item *i* under time point *t* where λ_{imt} is collected into a weight matrix Λ. A major advantage of the SLTM involves a particular processing ability θ_{jm} dependent on λ_{imt}. The SLTM is useful for data where items do not have the same discrimination on ability.

Tavares and Andrade (2006) proposed the 3PL LIRT model, and this model can be seen as an extension of Andersen's (1985) model. Tavares and Andrade's model assumed that the item parameters are known and fixed over time points; however, latent ability parameters describe the changes over time points. The 3PL LIRT model can be written as (Tavares & Andrade, 2006)

$$P\left(Y_{ijt}=1|\,\theta^*_{jt},a_i,b_i,c_i\right)=c_i+\left(1-c_i\right)\frac{\exp\left[a_i\left(\theta^*_{jt}-b_i\right)\right]}{1+\exp\left[a_i\left(\theta^*_{jt}-b_i\right)\right]}, \tag{59}$$

where $P\left(Y_{ijt}=1|\,\theta^*_{jt},a_i,b_i,c_i\right)$ is the probability of a correct response from respondent *j* on item *i* at time point *t* given the respondent ability θ^*_{jt}, item discrimination a_i, item difficulty b_i, and lower asymptote c_i. θ_j is $\left(\theta_{j1},\ldots,\theta_{jt}\right)^T \sim \mathrm{MN}_t(\mu,\Sigma)$, where $\mathrm{MN}_t(\mu,\Sigma)$ is the *t*-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ (von Davier et al., 2011).

In addition to the LIRT stated above, McArdle and Nesselroade (1994) proposed a latent change score (LCS), known as latent difference score (LDS), analysis.

2.3.3 Latent Change Score (LCS)

The LCS model is conducted within the framework of structural equation modeling (SEM) of longitudinal data. The LCS combined the concept of the structural modeling with latent growth modeling (McArdle, 2001). It is parameterized change as a function of the traditional growth terms, including intercept, slope, quadratic terms, and proportional growth. Proportional growth means that the change from one variable to another variable is dependent on the level at the previous variable, and it can be constrained to the same value across time or can differ in value across time (Keller & EL-Sheikh, 2011). McArdle and Hamagami (2004) suggested that the key features of the LCS model can straightforwardly define "changes as an accumulation of the first differences among latent variables" (p. 314) but not directly define the weights of the shape or timing of the change over time for the group (Λ_{jt}). In addition, the LCS includes other features (Keller & EL-Sheikh, 2011). First, the LCS can model within- and between- individual variance. Second, it can handle missing data by using full information maximum likelihood (FIML) estimation. Third, the assumption of sphericity or compound symmetry does not exist in this model. Fourth, this model can predict an additional aspect of change of individual differences by the parameterizing change. Fifth, this model includes simultaneously all statistical components of change.

Generally, researchers assume to measure a homogeneous sample of respondent *j* independently drawn from the population of interest. Assuming that we have measured the same repeatedly observed score (Y_{jt}) on at least two occasions, that is $T > 1$. In addition, the

measurement error is dealt with based on classical test theory. The equation of the observed score at any occasion *t* can be written as (McArdle, 2001)

$$Y_{jt} = y_{jt} + \varepsilon_{jt}, \tag{60}$$

where Y_{jt} is the observed score on a variable *y* for respondent *j* at occasion *t*, y_{jt} is the score of a latent variable *y* for respondent *j* at occasion *t*, and ε_{jt} is the error scores of a variable *y* for respondent *j* at occasion *t*. It is assumed that the scores of measurement error have a zero mean (μ_ε), have a non-zero variance $\left(\sigma_\varepsilon^2\right)$ and the same variance at each occasion, and are independent of any other component in the model (McArdle, 2001; McArdle & Hamagami, 2001). The latent variable is modeled as a function of the latent variable on the previous occasion and the degree of change, that is, latent change score.

A latent change score Δy_{jt} of any two latent or observed scores can be defined as (Ghisletta & McArdle, 2012; Keller & EL-Sheikh, 2011; McArdle, 2009; McArdle & Hamagami, 2001; McArdle & Nesselroade, 1994)

$$\Delta y_{jt} = y_{jt} - y_{j(t-\Delta t)}, \tag{61}$$

or

$$y_{jt} = y_{j(t-\Delta t)} + \Delta y_{jt}, \tag{62}$$

and the equation of the observed variables can be defined as

$$Y_{jt} = \eta_1 + \sum_{t=2}^{T} \Delta y_{jt} + \varepsilon_{jt}, \tag{63}$$

where Δt is the interval of time, and it is known and fixed; normally Δt is fixed to1. Thus, $y_{j(t-\Delta t)} = y_{j(t-1)}$. Δy_{jt} is the difference between *y* at the current occasion *t* and *y* at the prior occasion (*t*-1); y_{jt} is the score of the variable *y* for respondent *j* at the current occasion *t*; and $y_{j(t-\Delta t)} = y_{j(t-1)}$ is the

score of the variable *y* for respondent *j* at a prior occasion (*t*-1). Since the latent score is unconnected from the model based error scores, the latent change score is different from an observed change score (McArdle & Hamagami, 2004). In other words, the latent change scores Δy_{jt} emphasize the change in a variable *y* from previous occasion $t - 1$ to current occasion *t*, that is, the latent change score Δy_{jt} is computed by a current observed value y_{jt}, and subtracts a previous observed value $y_{j(t\text{-}1)}$ (Keller & EL-Sheikh, 2011; McArdle & Hamagami, 2001; Newson, 2015). This means that the score of the variable *y* for the respondent *j* at the current occasion Y_{jt} is formed as the unit-weighted sum of the latent score at the previous occasion $y_{j(t\text{-}1)}$ plus the latent change score Δy_{jt} for respondent *j* from the previous occasion *t*-1 to the current occasion *t* shown in Figure 8 (Newson, 2015).

The assumptions of the LCS include "the separation of individuals scores from group parameter," (McArdle & Hamagami, 2001, p. 150) a constant time interval, that is change time is equal to 1, and the separation of the latent score (y_t) from the error scores (ε_t) (McArdle & Hamagami, 2001). There are several models related to latent change scores, including no change score model (NCSM), linear change score model (LCSM), dual change score model (DCSM), and triple change score model (TCSM).

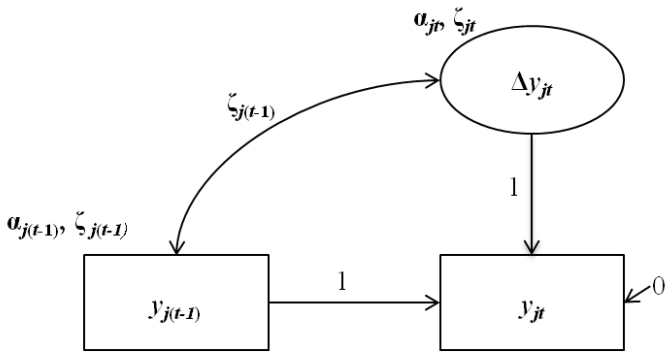

Figure 8. Latent change score model for two-time points.

First, the NCSM is the simple baseline model. This model does not have latent changes to capture the differences between occasions; however, this model allows systematic individual differences and random error at all time points. In other words, the NCSM only includes the latent intercept η_1, which is the unit constant, with a variance ψ_{11} and mean $\mu_{\eta 1}$. The $\mu_{\eta 1}$ represents fundamental distribution to all observed variables via the set of fixed unit regressions to all manifested scores (McArdle & Nesselroade, 2014). The NCSM can be written as (McArdle & Grimm, 2010)

$$\begin{aligned} \Delta y_{jt} &= 0, \\ Y_{jt} &= \eta_{1j} + \varepsilon_{jt}, \end{aligned} \tag{64}$$

where Δy_{jt} is the change score for respondent j at occasion t, Y_{jt} is the observed scores for respondent j at occasion t, η_{1j} is the initial level for respondent j, and ε_{jt} is the random error for respondent j at occasion t.

Second, the LCSM is essentially extended from the NCSM, but it includes parameters of a latent slope η_2 with a mean $\mu_{\eta 2}$, a variance ψ_{22}, and a covariance ψ_{12} on the latent change scores shown in Figure 9. Since the mean slope of this model represents "the average amount of change per unit of time" (McArdle et al., 2014, p. 436), this model is also called a constant change model. The coefficient of this model is unified from the latent slope to each latent change score. The LCSM is the same as what a researcher wants to obtain with any linear change model. In other words, the LCSM is the same as the linear latent growth curve model (McArdle & Nesselroade, 2014). The LCSM can be written as (McArdle & Grimm, 2010)

$$\begin{aligned} \Delta y_{jt} &= \eta_{2j}, \\ Y_{jt} &= \eta_{1j} + \left(\sum_{t=2}^{T} \eta_{2j} \right) + \varepsilon_{jt}, \end{aligned} \tag{65}$$

where η_{2j} is the latent slope of the model and $\sum_{t=2}^{T}\eta_{2j}$ is the sum of the latent change score up to time *t*. Y_{jt} can be expressed based on Figure 9 as (McArdle & Grimm, 2010)

$$\begin{aligned} Y_1 &= \eta_{1j} + \varepsilon_{1j}, \\ Y_2 &= \eta_{1j} + \eta_{2j} + \varepsilon_{2j}, \\ Y_3 &= \eta_{1j} + \eta_{2j} + \eta_{2j} + \varepsilon_{3j}, \end{aligned} \tag{66}$$

or we can simply write this as

$$Y_{jt} = \eta_{1j} + \eta_{2j(t-1)} + \varepsilon_{jt}. \tag{67}$$

According to Figure 9, circle and ellipsis represent latent variables, rectangular represents an observed variable, and the triangle is a constant score. The latent intercept (η_1) only affects the first observed variable (Y_1), but the second observed variable (Y_2) is affected by the latent

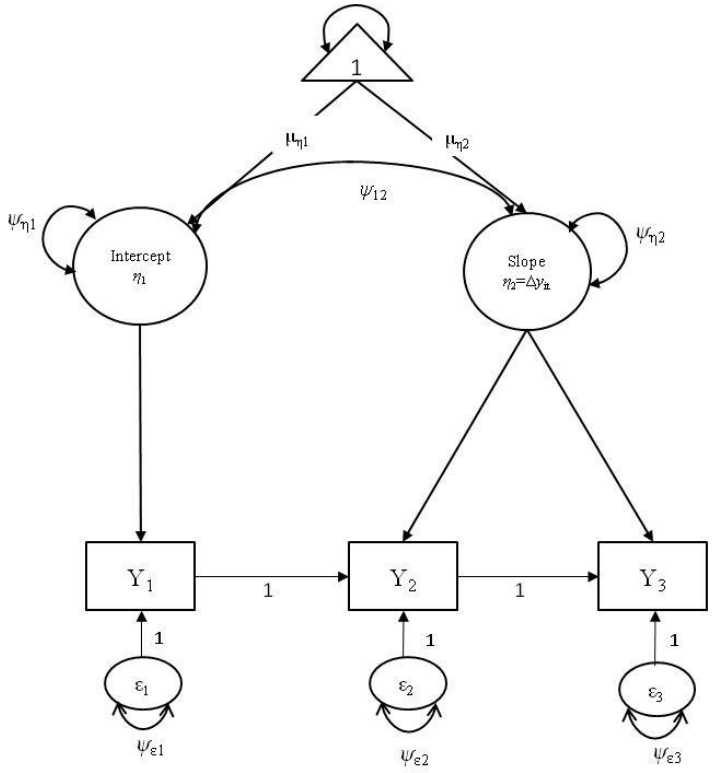

Figure 9. The linear change score model path diagram. (Note that the path diagram is modified by McArdle & Nesselroade, 2014.)

slope and the first observed variable. Since all coefficients are a fixed value of 1pointing ahead in time, any change of the second observed variable is conveyed to the third observed variable (Y_3). Thus, we can interpret that this is an accumulation of effects (McArdle & Nesselroade, 2014). For the LCSM, the mean of latent slope describes the mean of change observed from *t*-1 to *t*, and the variance of latent slope allows changes in individual difference. The mean, variance, and covariate are free to estimate in the LCSM.

Third, the change scores of DCSM directly include proportional change parameter β to the previous latent change score shown in Figure 10 (McArdle & Grimm, 2010). This model allows changes in the common factor scores, although the common factors are defined as invariant across times. The DCSM can be written as (McArdle & Grimm, 2010)

$$\Delta y_{jt} = \beta \bullet Y_{j(t-1)}, \tag{68}$$

$$Y_{jt} = \eta_{1j} + \left(\sum_{t=2}^{T} \eta_2 \alpha + \beta \bullet Y_{j(t-1)} \right) + \varepsilon_{jt}, \tag{69}$$

where Δy_{jt} is the change score from a previously observed variable, the coefficients of α and β describe the change. α is an additive change parameter that is a shift in the change at each occasion (McArdle & Hamagami, 2004), and β is proportional change parameters between the prior occasion y_t and current occasion $y_{(t\text{-}1)}$ that describe how each variable affects others across time points (McArdle & Grimm, 2010). Thus, four-time points can be defined as

$$\begin{aligned} Y_1 &= \eta_1 + \varepsilon_1, \\ Y_2 &= \eta_1 + \left(\eta_2 + \beta \bullet y_1 \right) + \varepsilon_2, \\ Y_3 &= \eta_1 + \left(\eta_2 + \beta \bullet y_1 + \eta_2 + \beta \bullet y_2 \right) + \varepsilon_3, \\ Y_4 &= \eta_1 + \left(\eta_2 + \beta \bullet y_1 + \eta_2 + \beta \bullet y_2 + \eta_2 + \beta \bullet y_3 \right) + \varepsilon_4, \end{aligned} \tag{70}$$

or we can generally write

$$Y_{jt} = \eta_{1j} + \eta_{2j(t-1)} + \left(\sum_{t=2}^{T} \beta \bullet y_{(t-1)} \right) + \varepsilon_{jt}. \quad (71)$$

Fourth, the TCSM can estimate the additive, proportional, and latent basis coefficients shown in Figure 11 (McArdle & Nesselroade, 2014). The latent basis coefficient of this approach is not a fixed value of 1 pointing ahead in time. Rather, it is free to estimate the latent basis coefficients of change (α_{t2}) (McArdle & Nesselroade, 2014). For instance, when the weight for the first time point is set at α_{12}=0, the weight of change score from first time point (Y_1) to second time point (Y_2) is set at α_{22} = 1, and the weight of the change score from second time point (Y_2) to third time point (Y_3) α_{32} is freely estimated. When these weights are estimated at unity, the linear curve is the optimal curve. The TCSM can be written as (McArdle & Grimm, 2010)

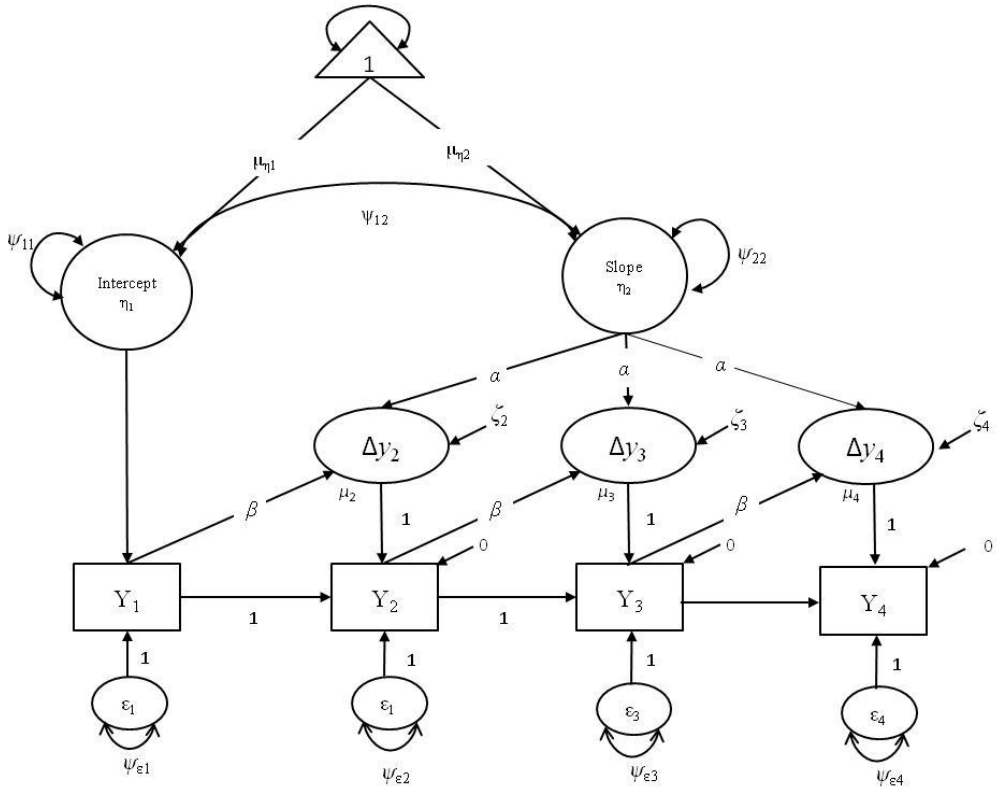

Figure 10. The dual change score model path diagram. (Note that the path diagram is modified by McArdle & Nesselroade, 2014.)

$$\Delta y_{jt} = \alpha_t \eta_{2j} + \beta \bullet y_{j(t-1)} \ , \tag{72}$$

$$Y_{jt} = \eta_{1j} + \left(\sum_{t=2}^{T} \alpha_t \eta_{2j} + \beta \bullet y_{j(t-1)} \right) + \varepsilon_{jt}. \tag{73}$$

Thus, four-time points can be defined as

$$\begin{aligned} Y_1 &= \eta_1 + \varepsilon_1, \\ Y_2 &= \eta_1 + \left(\alpha_2 \eta_2 + \beta \bullet y_1 \right) + \varepsilon_2, \\ Y_3 &= \eta_1 + \left(\alpha_2 \eta_2 + \beta \bullet y_1 + \alpha_3 \eta_2 + \beta \bullet y_2 \right) + \varepsilon_3, \\ Y_4 &= \eta_1 + \left(\alpha_2 \eta_2 + \beta \bullet y_1 + \alpha_3 \eta_2 + \beta \bullet y_2 + \alpha_4 \eta_2 + \beta \bullet y_3 \right) + \varepsilon_4, \end{aligned} \tag{74}$$

or we can generally write

$$Y_{jt} = \eta_{1j} + \alpha_t \eta_{2j(t-1)} + \left(\sum_{t=2}^{T} \beta \bullet y_{(t-1)} \right) + \varepsilon_{jt} \, . \tag{75}$$

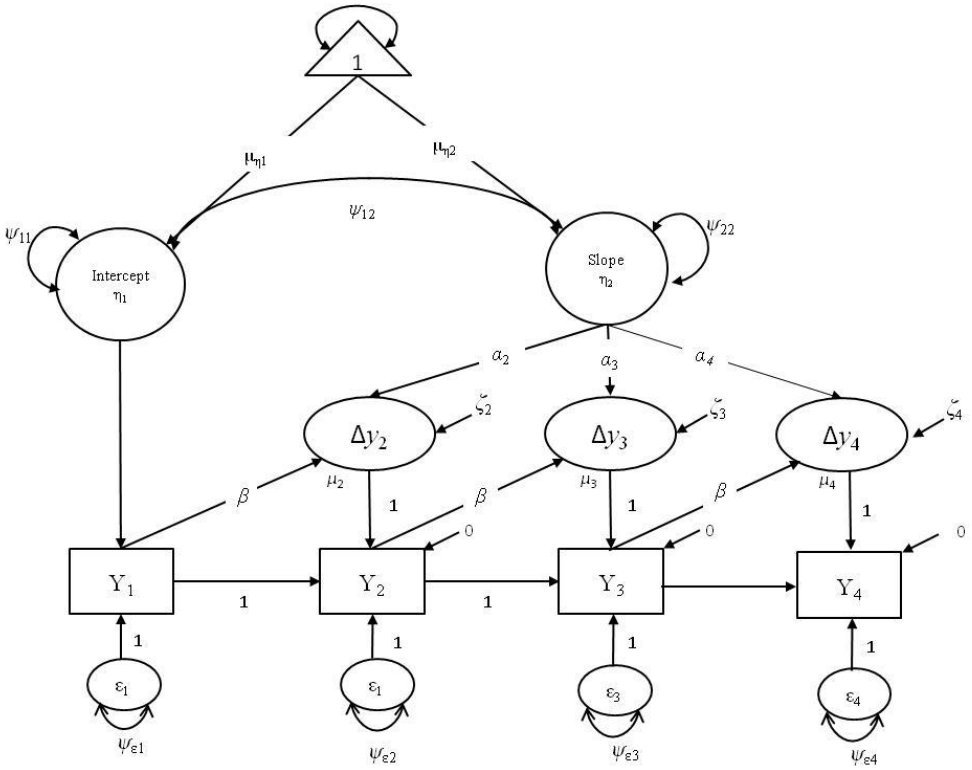

Figure 11. The triple change score model path diagram. (Note that the path diagram is modified by McArdle & Nesselroade, 2014.)

The latent change score also can be extended to many other forms of change models, such as auto-regressive change score (ACS) model (McArdle, 2001), free change score (FCS) model (McArdle et al., 2014) and change scores based on second-order difference operators (Hamagami & McArdle, 2007).

2.4 The Basic Concept of Latent Growth Modeling (LGM)

Traditional methods, such as univariate- and multivariate- analysis of variance/covariance, auto-regressive, and multiple regression methods, are used to study longitudinal data. However, the limitation of these methods was that they only analyzed the mean changes and treated differences among individuals as error variance; some error variance might contain helpful information related to change (Curran, Obeidat, & Losardo, 2010). In addition, those methods both include variations of within-subject and between-subject. Nonetheless, these two kinds of variations are not the same; thus, they should be analyzed differently. For instance, the between-subject variations result in dependencies among observed variables (Takane & de Leeuw, 1987).

Researchers have extended a statistical approach, growth curve model, to describe individual differences and the nature of growth across time. Growth curve model is a flexible method to model change over time, it allows for exploring linear and nonlinear trends and change in individual differences, and it allows estimating between-person differences in within-person change over time (Curran et al., 2010; Newsom, 2015). There are two popular methods of growth curve model for analyzing longitudinal data. First, growth curve analysis, which is based on multilevel modeling (MLM), known as hierarchical linear modeling (HLM), allows for dealing with data containing multilevel structure. Second, latent growth modeling (LGM), also known as latent growth curve model (Preacher et al., 2008) or latent curve analysis (Meredith &

Tisak, 1990), offers additional flexibility and integrates growth curve model in structural equation modeling (SEM) (Newsom, 2015). Since the growth curve model in this study is based on SEM framework, SEM will be briefly discussed in the next section.

2.4.1 Structural Equation Modeling (SEM)

SEM can be viewed as a combination of ANOVA, confirmatory factor analysis (CFA), multiple regression, and path analysis, and it is widely used in the behavioral and social sciences fields. SEM is a generalization of linear modeling providing that mathematical and statistical devices and testing hypothesized patterns of complex relationships between observed (measured) variables, which denotes indicators, and unobserved (latent) variables, which indicates a construct, cannot be directly observed (Preacher et al., 2008). Researchers are interested in latent variables (factors), and the relationships between the latent variables are symbolized by the factor loadings between those variables. SEM not only can deal with regression, including single or multiple linear regressions, but also can deal with a system of regression equations; thus, it is a flexible model. The main strength of SEM is to compare the model to the data; that is, the model implies that the structure of the means and covariances can be compared to the means and covariances of the data. This comparison, known as fit-statistics, evaluates the matching of model and data. The second advantage of SEM is that direct and indirect effect between variables can be estimated. SEM can also be useful in several ways (McArdle & Bell, 2000). First, it organizes concepts related to data analysis into scientific models. Second, it provides instruments for the estimation of mathematical components and for the assessment of the statistical features of models. Third, it allows a flexible method to deal with incomplete data sets. Fourth, it contains flexible provisions for models with latent variables. Overall, there are two

important purposes in SEM. First is obtaining an estimate of the model parameter. Second is assessing the fit of the model. Although SEM has several advantages and is very useful, the required large sample sizes are one of its limitations.

SEM includes a structural model and measurement model. The structural model denotes the casual relationships between latent variables and accounts for the casual effects. For continuous indicators, the structural model can be defined as (Muthén, 1983)

$$\eta = \pi + \mathrm{B}\eta + \varepsilon, \tag{76}$$

where η is a latent dependent variable; π is an intercept , normally fixed to zero; B is the coefficient matrix of latent dependent variable η; and ε is a structural equation error .

The measurement model defines the relationship between a latent variable and its observed indicators (Jöreskog & Sorbon, 1996), and the equation can be written as

$$y_{ij} = \pi_i + \Lambda_i \eta_j + \delta_{ij}, \tag{77}$$

where y_{ij} is an observed indicator variable, i =1,…, n; π_i is an intercept for measurement, normally fixed to zero; η_j is a latent variable; Λ_i are factor loadings about the relations of the latent response variable (y_{ij}) with the latent variables (η_j); and δ_{ij} are measurement errors, and it follows the normal distribution; that is $\delta \sim N(0,1)$. Simultaneously, SEM can be visualized by path diagram, which was invented by Sewall Wright (1921), shown in Figure 12.

In Figure 12, two latent variables ($F1$ and $F2$), represented by a circle, cannot estimate directly. The observed variables (Y_1 through Y_6), represented by rectangles, estimate the latent variables. Error terms (ε_1 through ε_6) are represented by ellipses. Single-headed arrows indicate causal relationships, and double-headed arrows define correlation between two latent factors (i.e., covariances ψ_{12}) and also represent the variance of latent variables. The two latent factors

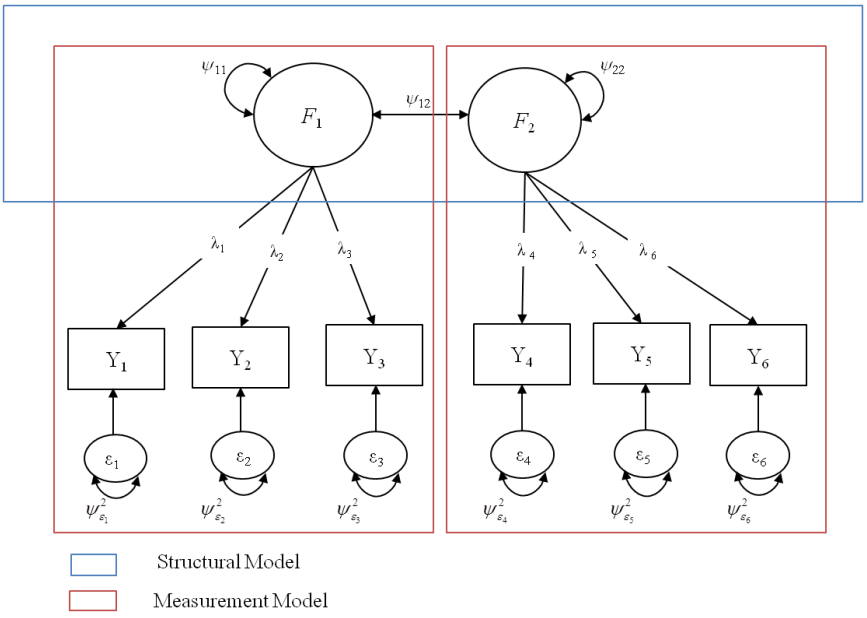

Figure 12. SEM model with two common factors path diagram.

have their own variances, ψ_{11} and ψ_{22}, respectively, and they produce variances $\psi^2_{\varepsilon_t}$ in each observed variable (i.e., Y_1-Y_6) through the factor loadings (λ_1 to λ_6) (McArdle & Nesselroade, 2014). Each observed variable has a unique $\psi^2_{\varepsilon_t}$, and assuming the unique $\psi^2_{\varepsilon_t}$ is independent of other unique and common factors. Normally, the path diagram does not show the intercept π_i because of setting to zero.

In SEM, the parameter estimation and computing model fit are achieved with MLE using computer programs, such as LISREL, EQS, AMOS, and M*p*lus, assuming observed variables to be multivariate normal distribution when the observed data are continuous and recommending the minimum sample sizes be at least 200 (Kline, 1998). When the sample sizes are large enough, the statistical tests are almost significant; thus, the model will be rejected, although the data are described very well by the model. Because of the sensitivity of the statistical test for sample sizes, we should employ alternative fit indices, that is goodness-of-fit, to evaluate model

fit. The goodness-of-fit indices include goodness-of-fit (GFI), Tucker-Lewis index (TLI), and non-normed fit index (NNFI); when these fit indices have the value 1, the model is fitted perfectly. The rule of thumb for accepting the model is a value of at least .90, and judging the model fit as good requires a value of at least .95. In addition to the fit indices stated above, root mean square error of approximation (RMSEA) can also be used to judge the model fit as good. When RMSEA is small, which is less than .05, the approximation is good (Hox & Bechger, 1998).

In addition to continuous indicators, SEM has been developed to handle dichotomized variables since the mid-1970's. Dichotomous indicators are crucial to develop because observed variables are dichotomous with non-equidistant scale steps, especially in the social and behavioral sciences. For dichotomous indicators, the structural model is described the same as the continuous indicators in Equation 76. Here y_i^* is used to model categorical outcomes of the measurement model (Muthén, 1983):

$$y_{ij}^* = \pi_i + \Lambda_i \eta_j + \delta_{ij} . \tag{78}$$

Equation 78 is identical to Equation 77 except y_{ij}^* is a continuous latent response variable of respondent *j* on item *i* related to an unobserved variable; Λ_i is the factor loadings and contains matrices of fixed parameters. δ_{ij} is measurement errors that are uncorrelated to the latent variables η_j and each other. Because y_{ij}^* is unobserved, the mean for single population can be set to zero, and its variances are arbitrary, normally setting variance to 1 for convenience (Muthén, 1983). In Equation 78, a continuous latent response variable y_{ij}^* is introduced to represent the propensity of the occurrence of a certain category in a binary outcome. In that case, a binary outcome is considered an observed categorical indicator of an unobserved continuous latent

response variable (Muthén, 1979, 1984). For a binary outcome, a continuous latent response variable y_{ij}^{*} is related to the binary outcome of the observed response y_i via a threshold τ_i as in the following:

$$y_i = \begin{cases} 1 \text{ if } y_i^* \geq \tau_i \\ 0 \text{ if } y_i^* \leq \tau_i \end{cases}, \tag{79}$$

where the threshold τ_i is usually assumed to be zero because of an identification purpose. Since y_i^* can be considered as continuous latent response variables underlying observed indicators y_i, y_i are assumed to reflect the y_i^*, and y_i^* is usually assumed to be normal.

When the observed data are categorical to follow the normal distribution, IRT and SEM, in fact, cover the same types of categorical data. Therefore, there is a special relationship between IRT and SEM methods. Indeed, Takane and de Leeuw (1987) have proved that these two models are equivalent. The following section will briefly discuss the connection of IRT and SEM for categorical data.

2.4.1.1 IRT-SEM Model

When observed variables (y_i) are categorical, assuming them to be binary (i.e., y_i =1 or y_i =0), observed variables assume reasonably that they are Bernoulli random variables. Let $y = (y_1, \ldots, y_n)$, the random vector of item response pattern; θ is ability, which is an m-component random vector and cannot be measured directly, and its density function is defined by $g(\theta)$. The domain of θ is denoted by Θ, which is the multidimensional region, ranging from $-\infty$ to ∞, and θ assumes to follow a multivariate normal distribution with mean 0 and identity matrix; that is $\theta \sim N(0, (I))$. The two-parameter IRT model can be chosen to be either normal ogive or logistic

models. The two-parameter normal ogive model in IRT is defined (Bock & Aitkin, 1981; Takane & de Leeuw, 1987)

$$P(Y = y) = \int_{\Theta} P(Y = y \mid \theta) g(\theta) d\theta \,, \tag{80}$$

where $P(Y = y \mid \theta)$ indicates the conditional probability of observing item response pattern *y* given θ. Under local independence, $P(Y = y \mid \theta)$ can be defined as

$$P(Y = y \mid \theta) = \prod_{i=1}^{n} \left[P_i(\theta) \right]^{y_i} \left[1 - P_i(\theta) \right]^{1-y_i} \,, \tag{81}$$

where $P_i(\theta)$ gives the probability that a correct response ($Y_i = 1$, $i = 1, 2, \ldots, n$) will take place for a respondent whose ability is given by θ, and $P_i(\theta)$ is given as

$$P_i(\theta) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(\alpha_i \theta + \delta_i), \tag{82}$$

where $z_i = \alpha_i \theta + \delta_i$ and $\delta_i = -\alpha_i \beta_i$. Note that the parameter in two-parameter normal ogive model in IRT is usually reported in the form of $\Phi \alpha_i (\theta - \beta_i)$ where α_i is the parameter of discrimination; and β_i is the parameter of difficulty; Φ indicates the cumulative distribution function of the standard normal distribution. For the two-parameter logistic model, the form of $\Phi \alpha_i (\theta - \beta_i)$ can be defined as

$$\Phi \alpha_i (\theta - \beta_i) \approx \frac{\exp\left[1.7 \alpha_i (\theta_j - \beta_i) \right]}{1 + \exp\left[1.7 \alpha_i (\theta_j - \beta_i) \right]} \,, \tag{83}$$

where the value of 1.7 is merely stated that obtained parameter of discrimination from the logistic model is about 1.7 times as large as the equivalent parameters of the normal ogive models (Glockner-Rist & Hoijtink, 2003).

When the mean is assumed to be zero and the variance to be one, the SEM for binary outcomes can be defined as

$$P(Y = y) = \int_U h(y)dy,$$

$$= \int_U \left(\int_R f(y \mid w)g(w)dw \right) dy,$$

$$= \int_R g(w) \left(\int_U f(y \mid w)dy \right) dw, \tag{84}$$

where U is the multidimensional region, ranging from $-\infty$ to ∞. U is defined by the direct product of intervals, which makes if $y_i = 1$, $U_i = (u_i, \infty)$ and if $y_i = 0$, $U_i = (-\infty, u_i)$. $f(y \mid w)$ indicates the conditional density of y given w. Note that $y \mid w \sim N(Dw, Q^2)$, and Equation 84 involves no distribution assumption; the local independence assumption is not required. Therefore, the $\int_U f(y \mid w)dy$ can be defined as

$$\int_U f(y \mid w)dy = \prod_{i=1}^{n} \int_{u_i} f_i(y_i \mid w)dy_i,$$

$$= \prod_{i=1}^{n} \left(\int_{u_i}^{\infty} f_i(y_i \mid w)dy_i \right)^{y_i} \left(1 - \int_{u_i}^{\infty} f_i(y_i \mid w)dy_i \right)^{1-y_i}, \tag{85}$$

where

$$\int_{u_i}^{\infty} f_i(y_i \mid w)dy_i = \Phi\left(\frac{d_i w - u_i}{q_i} \right), \tag{86}$$

where q_i is the i-th diagonal element of $\text{Var}(Q^2)^{1/2}$. If

$$\alpha_i = \frac{d_i}{q_i} \tag{87}$$

and

$$\delta_i = \frac{-u_i}{q_i}, \tag{88}$$

Equation 86 becomes

$$\Phi\left(\frac{d_i w - u_i}{q_i}\right) = \Phi(\alpha_i w + \delta_i). \tag{89}$$

As alluded to above, since $\delta_i = -\alpha_i \beta_i$, $\Phi(\alpha_i w + \delta_i) \approx \Phi\alpha_i\left(w - \beta_i\right)$. Kamata and Bauer (2008) provided the general transformation equations to convert item parameter estimates in the SEM

$$\alpha_i = \frac{1.7\Lambda_i \, \mathrm{Var}(\eta)^{1/2}}{\mathrm{Var}(\delta_{ij})^{1/2}}, \tag{90}$$

and

$$\beta_i = \frac{\tau_i - \Lambda_i E(\eta)}{\Lambda_i \, \mathrm{Var}(\eta)^{1/2}}, \tag{91}$$

where η is the latent variable; Λ_i and τ_i are the factor loading and threshold for item *i*, respectively; $E(\eta)$ and $\mathrm{Var}(\eta)$ are the mean and variance of the latent variables; and $\mathrm{Var}(\delta_{ij})$ is the residual variance of item *i*. Thus, Equation 89 is equivalent to Equation 82 (Kamata & Bauer, 2008; Takane & de Leeuw, 1987).

In sum, the ability (i.e., latent trait) in IRT is assumed to be unidimensional, whereas the latent variable (i.e., ability in IRT) in SEM tends to be multidimensional. However, IRT and SEM are closely related when items are categorical data. As we know, latent variables cannot be directly measured; they are measured through a set of items. When the latent variables in SEM are unidimensional, the model is the same as a graded item response model with a probit link (Titman, Lancaster, & Colver, 2016). Takane and de Leeuw (1987) showed dichotomous variables of SEM to be equivalent to the two-parameter normal ogive model in IRT. Note that

the connection between the IRT model and SEM has been discussed in recent research (Glockner-Rist & Hoijtink, 2003; Kamata & Bauer, 2008; Titman et al., 2016).

2.4.2 Latent Growth Modeling (LGM)

Latent growth modeling (LGM) is commonly used in modeling the growth of the latent construct measured by a scale that contains a set of items. It describes the relationship between the repeated measurement of the same observed variable and the metric of occasions. In addition, LGM makes it possible to explore the relationship between latent predictor variables in change, measure the effects of change on other factors, develop better hypothesis articulation, construct parallel growth curves, and provide greater statistical power (Newsom, 2015; Preacher et al., 2008). Furthermore, it directly allows investigation of *intra*individual (within-person) change over time and *inter*individual (between-person) variability in intraindividual change (Preacher et al., 2008). Basically, LGM delineates individuals' behavior in terms of initial levels and their growth from or to those levels. LGM decides the variability across individuals in both initial levels and slope, and it offers an approach to inspecting the contribution of other variables (or constructs) to account for those initial levels and slope (Hancock, Kuo, & Lawrence, 2001).

LGM was originally developed by both Tucker (1958) and Rao (1958); William Meredith was the first to relate it to structural equation modeling (SEM) in 1985 (McArdle & Bell, 2000; Meredith & Tisak, 1990). LGM employs a concept derived from the confirmatory factor analysis (CFA) as a special case of SEM applied to longitudinal data (Preacher et al., 2008; McArdle & Bell, 2000). The simple linear LGM includes two latent variables, η_1 representing an intercept and η_2 representing a slope, and observed variables, U_1 to U_3, as shown in Figure 13. The latent variable indicates that constructs cannot be directly measured, but it can

be measured by observed variables. In LGM, the relationship between latent variables and indicators is similar to the CFA model. The effects of latent variables on their indicators are called loadings, which describe trends over time in observed variables that are repeated measures of the same observed variable *U*, where *U* has been measured from time *t* (*t* =1, …, *T*) on respondent *j* (*j* =1, …, *N*), that is U_{jt} (Ghisletta & McArdle, 2012).

In Figure 13, the intercept η_1 is the level of outcome measure, and the slope η_2 is specified to denote shape of the growth curve of change over time (Preacher et al., 2008). LGM model requires at least one fixed value for each latent variable and one fixed zero to separate latent variables for estimation (McArdle & Nesselroade, 2014). For example, we fixed values of 1 for

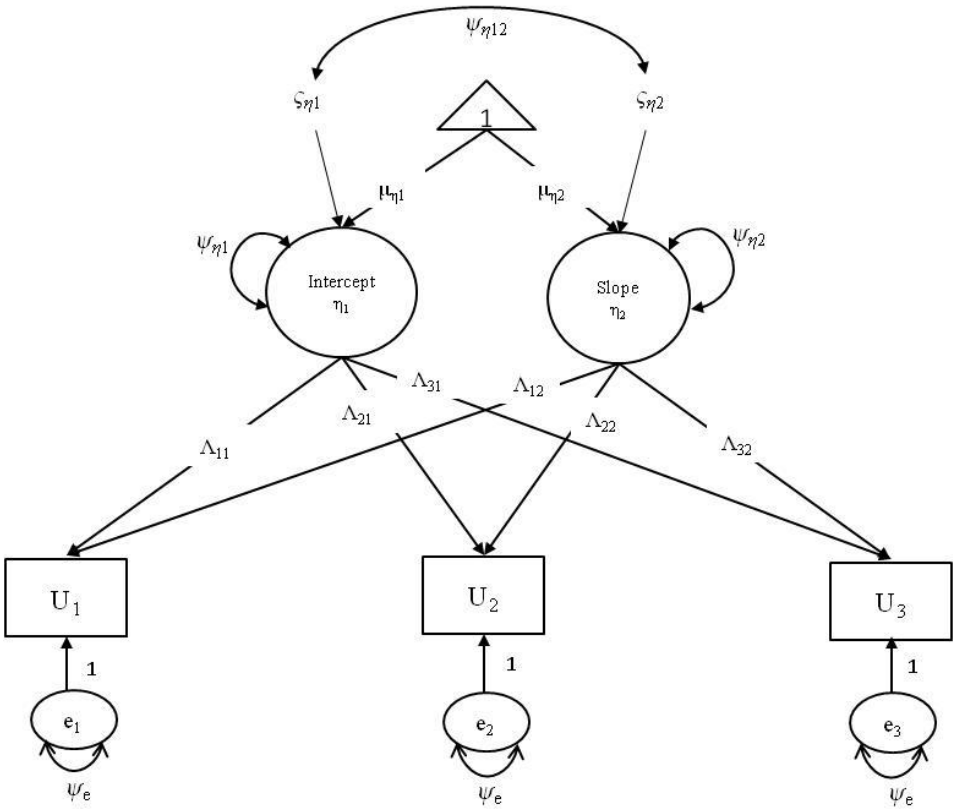

Figure 13. The linear latent growth models for three time points path diagram.

the intercept, and we fixed zero in the slope at the first time point in this study. A triangle means a constant of value of 1 that includes the means of the intercept and slope. Circles or ellipses indicate latent variables, and rectangles indicate observed variables. One-headed arrows indicate fixed or group coefficients (i.e., fixed effect) that are structural weights, and a double-headed arrow indicates random or individual's features (i.e., random effect). The intercept, η_1, which is constant for any given individual across time, represents the level of the outcome measure (i.e., observed variable) *U*. The loadings of intercept, Λ_{11} through Λ_{31}, are required to be equal to the unit value, normally fixed to 1 in order to represent the effects of a constant for longitudinal data.

The slope, η_2, represents the slope of an individual's change. The loadings of the slope, Λ_{12} through Λ_{32}, represent the shape of linearly increasing growth over time (McArdle & Nesselroade, 2014). The loadings of the slope are set based on a chosen time metric (Preacher et al., 2008; Duncan & Duncan, 2004). For instance, one employs a balanced set of basic weights as 0, 1, and 2, indicating the mean to reflect a one-unit change from one occasion to the next occasion. In addition to the balanced set of basic weights, an unbalanced set of basic weights as 0, 1, and 3 can be explained for the observed but unbalanced time delay between time points (McArdle & Nesselroade, 2014). $\mu_{\eta 1}$ and $\mu_{\eta 2}$ are the mean of the intercept and slope, respectively. The mean of the slope, $\mu_{\eta 2}$, indicates a linear increase or decrease across time. If the mean of the slope is significant, the increase or decrease across time in the level of the dependent variable is statistically different from 0 (Newsom, 2015). $\zeta_{\eta 1}$and $\zeta_{\eta 2}$ are residuals of the intercept and the slope, respectively. $\psi_{\eta 1}$ and $\psi_{\eta 2}$ are the variances of the intercept and the slope, respectively. The variance of the slope, $\psi_{\eta 2}$, indicates the individual differences in change. The significant variance of slope suggests that the slopes for individual cases are heterogeneous, whereas the insignificant variance of the slope indicates that individual cases are homogeneous

(Newsom, 2015). In addition, the intercept-slope covariance $\psi_{\eta 12}$ is able to estimate. U_1 through U_3 are equal spaces of observed variable U for longitudinal data. e_1 to e_3 are the error terms, which influence the interpretation of the model parameters by correcting the measured variances for the random error (Preacher et al., 2008).

The matrix algebra of LGM based on Figure 13 can be written as

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \\ \Lambda_{31} & \Lambda_{32} \end{bmatrix} \times \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}, \tag{92}$$

where the first column is τ, which is a linear function of intercepts, normally fixing to zero because of model identification and not shown in Figure 13, the second column is the loadings of the initial level (i.e., intercept), the third column is the loadings of the slope, the fourth column is the two latent variables, and the last column is the errors. The basis matrix notation can be written as

$$U = \tau + \Lambda\eta + e, \tag{93}$$

where U is the observed variables; τ is as stated above; η represents g latent variables that indicate the aspects of change; Λ, which indicates the function of time, is a matrix of loading of latent variables; and e is error terms. The measurement model indicates the relationship between the latent variables (η) and observed variables (U) for longitudinal data (Preacher et al., 2008). When $g = 2$, assuming the U variable measured at three equal-interval occasions, the equation of Figure 13 can be defined as:

$$U_{jt} = \eta_{j1}\Lambda_{t1} + \eta_{j2}\Lambda_{t2} + e_{jt}, \tag{94}$$

and a test for linear growth model could be conducted through the following loadings:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}. \tag{95}$$

The first column represents the true initial amount (i.e., intercept) Λ_{t1} of *U*, and the second column is the true rate of change (i.e., slope) Λ_{t2} across time points from that $\Lambda_{t1.}$ If Λ_{t1} is fixed to 1, then Equation 94 becomes

$$U_{jt} = \eta_{j1} + \eta_{j2}\Lambda_{t2} + e_{jt}, \tag{96}$$

where U_{jt} indicates the score on the observed variables for respondent *j* at time *t*; η_{j1} and η_{j2} are the intercept and slope for respondent *j*, respectively; Λ_{t2} is the loading of the slope, giving values, such as 0, 1, and 2, to represent time codes, and it determines the shape of the growth curve; and e_{tj} is a random normal error for respondent *j* on the observed variables at time *t* unexplained by the initial level and the rate of change (Ferrer, Balluerka, & Widaman, 2008). The η_{j1} and η_{j2} can be expressed as functions of latent means and residuals (i.e., individual deviations) away from latent means. The equation of η_{j1} and η_{j2} can be defined as

$$\eta_{j1} = \mu_{\eta_1} + \zeta_{\eta 1}, \tag{97}$$

$$\eta_{j2} = \mu_{\eta_2} + \zeta_{\eta 2}, \tag{98}$$

where $\mu_{\eta 1}$ and $\mu_{\eta 2}$ are the mean of intercept and slope, respectively, and $\zeta_{\eta 1}$ and $\zeta_{\eta 2}$ are the residuals of the intercept and slope, respectively. η_{j1} and η_{j2} are referred to as random coefficients, and $\zeta_{\eta 1}$ and $\zeta_{\eta 2}$ are referred to as random effects (Preacher et al., 2008). When substituting Equation 97 and 98 into Equation 96, this equation can be written as

$$U_{jt} = \mu_{\eta_1} + \mu_{\eta_2}\Lambda_{t2} + \zeta_{\eta 1} + \zeta_{\eta 2}\Lambda_{t2} + e_{jt}. \tag{99}$$

We also can derive a covariance structure based on Figure 13. The covariance structure indicates the variance and covariance of the population of observed variables for longitudinal data as a function of model parameters. The covariance structure is defined as

$$\Psi = \Lambda\zeta\Lambda' + \varepsilon_e, \tag{100}$$

where Ψ is the variance-covariance matrix of observed variable U ; Λ is the loading matrix of the latent variables; ζ is the variance-covariance matrix of the latent variables; and ε_e is the variance-covariance matrix of error terms, and it indicates that part of the variance in data is uncorrelated with the hypothesized latent curves (Bollen, 1989; Preacher et al., 2008).

In addition to deriving the covariance structure, the mean structure can also be derived. The mean structure indicates the population mean of observed variables for longitudinal data as a function of intercept and the mean of latent variables. The mean structure is defined as (Preacher et al., 2008)

$$\mu_\eta = \Lambda_\eta \Upsilon, \tag{101}$$

where μ_η is the matrix of population mean of observed variables; Υ is the means matrix of latent variables, (i.e., $\mu_{\eta 1}$ and $\mu_{\eta 2}$). As stated above, Λ_η is the matrix loading of the latent variables. Note that the mean structure, in fact, characterizes the population means of observed variable U as a function of τ and μ_η; however, τ normally is constrained to zero so that we can simplify measurement model and mean structure. The parameters of researchers interested in the LGM are included in the matrices of Λ, ζ, ε, and μ_η. Therefore, the simple linear LGM model of Figure 13 estimates a total of six parameters, including the means ($\mu_{\eta 1}$ and $\mu_{\eta 2}$) and variance and covariance ($\psi_{\eta 1}$, $\psi_{\eta 2}$, and $\psi_{\eta 1\, \eta 2}$) of latent variables and the constant errors (ψ_e).

2.4.3 Assumptions of LGM

There are several important assumptions in LGM with maximum likelihood estimation (MLE) (Bollen & Curran, 2006; Preacher et al., 2008). First, the means of the error term in Equation 93 and residuals in Equation 97 and 98 are assumed to be zero. This mean that if we measure the same person repeatedly at a given time point, we assume the means of error and residuals to be zero across given occasions. Second, the latent variables (i.e., the intercept and slope) are independent with the equation error term. Third, assuming the covariances of all variances within and between time points is zero, that is, the errors are independent over time. Fourth, assuming random intercepts and slopes are independent of other factors. Fifth, the errors of different individuals are independent.

2.4.4 The Strengths and Limitations of LGM

There are several advantages to using LGM for evaluating change. First, since LGM treats mean, variance, and covariance as random effects, LGM allows the estimation of an average growth trajectory (i.e., the mean of intercept and slope) of an individual and individual differences (i.e., the variance of intercept and slope) in change over time. Unlike repeated-measures ANOVA, which evaluates only mean growth patterns and treats variability as an error in a growth pattern, LGM estimates the growth trajectory and group means and variances of the growth factors separately for each individual. In addition, it allows the study of the predictors of those individual differences to answer questions related to which variables apply important effects on the rate of development (Hardy & Thiels, 2009; Duncan & Duncan, 2004).

Second, LGM has considerable flexibility because it is capable of investigating both linear and nonlinear change patterns, given at least two-time points. Two factors are adequate

and preferable for estimating linear growth, and more than two factors are able to investigate nonlinear, such as quadratic and cubic, growth (Duncan et al., 2006; Duncan & Duncan, 2004).

Third, LGM permits the incorporation of both time-varying and time-invariant covariates. Both time-invariant and time-varying variables can be included in models as predictors and outcomes of growth functions. Thus, it permits the researcher to explore the antecedents and consequences of development (Preacher et al., 2008; Duncan & Duncan, 2004; Duncan et al., 2006).

Fourth, LGM has the ability to assess the sufficiency of models by using model fit indices and model selection criteria and to explain measurement error by using latent repeated measures. LGM is able to explain some error in predictors and to investigate mediation hypotheses (Burchinal, Nelson, & Poe, 2006).

Fifth, for dealing with missing data, LGM is conducted by using the full information maximum likelihood (FIML) method, which is suggested to obtain maximum likelihood (ML), to estimate parameters. Unlike repeated-measures ANOVA, which either deletes missing data cases or imputes the value for missing data prior to the analysis, ML parameter estimation uses all available data. In other words, LGM takes all available information instead of deleting missing data cases. Thus, LGM is more useful for dealing with missing data than repeated-measures ANOVA (Hardy & Thiels, 2009; Preacher et al., 2008; Duncan et al., 2006).

Despite LGM possessing a number of benefits, it has some limitations. First, change is systematically associated with the passage of time based on a fundamental assumption. If change and the passage of time are not related, studying individual growth trajectories will not be very useful (Duncan & Duncan, 2004; Burchinal & Appelbaum, 1991). Second, LGM would not be appropriate for randomly varying within-subjects designs, varying within-person distributions of

time-varying covariates, and random missing data. Despite LGM having the ability to handle missing data and multilevel growth modeling, the analyses do not yet permit the flexibility of the random coefficient method (Duncan & Duncan, 2004; Duncan et al., 2006). Third, it relies on time-structured data and a slightly reduced power (Burchinal et al., 2006). Fourth, it lacks a mechanism for evaluating longitudinal measurement invariance (Sayer & Cumsille, 2001).

In order to have rigorous basis for meaningful scaling for investigating growth, LGM can be considered to incorporate with other approaches, such as IRT. This study employs the LGM to integrate the IRT approach with longitudinal data, that is, longitudinal item response theory – latent growth modeling (LIRT-LGM). The next section will discuss LIRT-LGM model.

2.5 Longitudinal Item Response Theory – Latent Growth Model (LIRT-LGM)

The LGM employs a single composite, which is normally summed or averaged over items to create a score. It is this score that is used to investigate change. The LGM model is directly fit to the vector of means and the matrix of covariances among single observed scores measured, that is, one score per individual, at each measurement occasion. As noted above, the LGM does not incorporate measurement errors of the indicators into the composite score. In order to overcome the limitations of the LGM, in the 1980s the LGM model was extended to incorporate multiple indicators. The model incorporating multiple indicators is known as the second-order latent growth modeling (Hancock et al., 2001; Sayer & Cumsille, 2001), curve-of-factors model (McArdle, 1988), latent variable longitudinal curve model (Tisak & Meredith, 1990), multivariate latent curve models (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997), or multiple indicator growth curve model (Chan, 1998). The manifest variables are to be used as the indicators of the latent variable (i.e., first-order factors) on each occasion. The growth

parameters (i.e., second-order factors) are used to explain variance in and covariance among the first-order factors and to investigate growth over time (Ferrer et al., 2008; Hancock et al., 2001; Sayer & Cumsille, 2001).

The basic path diagram of the second-order LGM is presented in Figure 14. The difference between the LGM and the second-order LGM is that the intercept η_1 and slope η_2 of the latent variables become the second-order factors. In addition, the first-order factors, F_1 to F_3, are included in the second-order LGM. These are the latent constructs measured by multiple indicators on each occasion. The loadings for the second-order factors are described in the first-order factors; the loadings of the intercept are set to 1 and for the slope are set based on the time metric (in this study, these are 0, 1, and 2). The mean and variance of the intercept and slope are the most interesting in applications as with the LGM. The loadings, λ, represent a regression slope associated with the observed score and the latent construct. The intercepts, τ, in the regression model are related to each indicator of the first-order factors. These values are set to zero in order that the means of the first-order can be identified constraints on the first-order factors (Newson, 2015). The equation in Figure 14 can be stated as

$$Y_{tj} = \tau_{ti} + \lambda_{ti} F_{tj} + \varepsilon_{tj}, \tag{102}$$

where

$$F_{tj} = \eta_1 + \Lambda_{2t}\eta_2 + \psi_{tj}, \tag{103}$$

$$\eta_1 = \mu_{\eta 1} + \zeta_{\eta 1}, \tag{104}$$

$$\eta_2 = \mu_{\eta 2} + \zeta_{\eta 2}, \tag{105}$$

where Y_{tj} is the score for respondent j at time t; τ_{ti} is the intercept of item i at time t, normally τ_{ti} is set to 0 and not shown on Figure 14; F_{tj}, is first-order factors for respondent j at time t; λ_{tj} is the

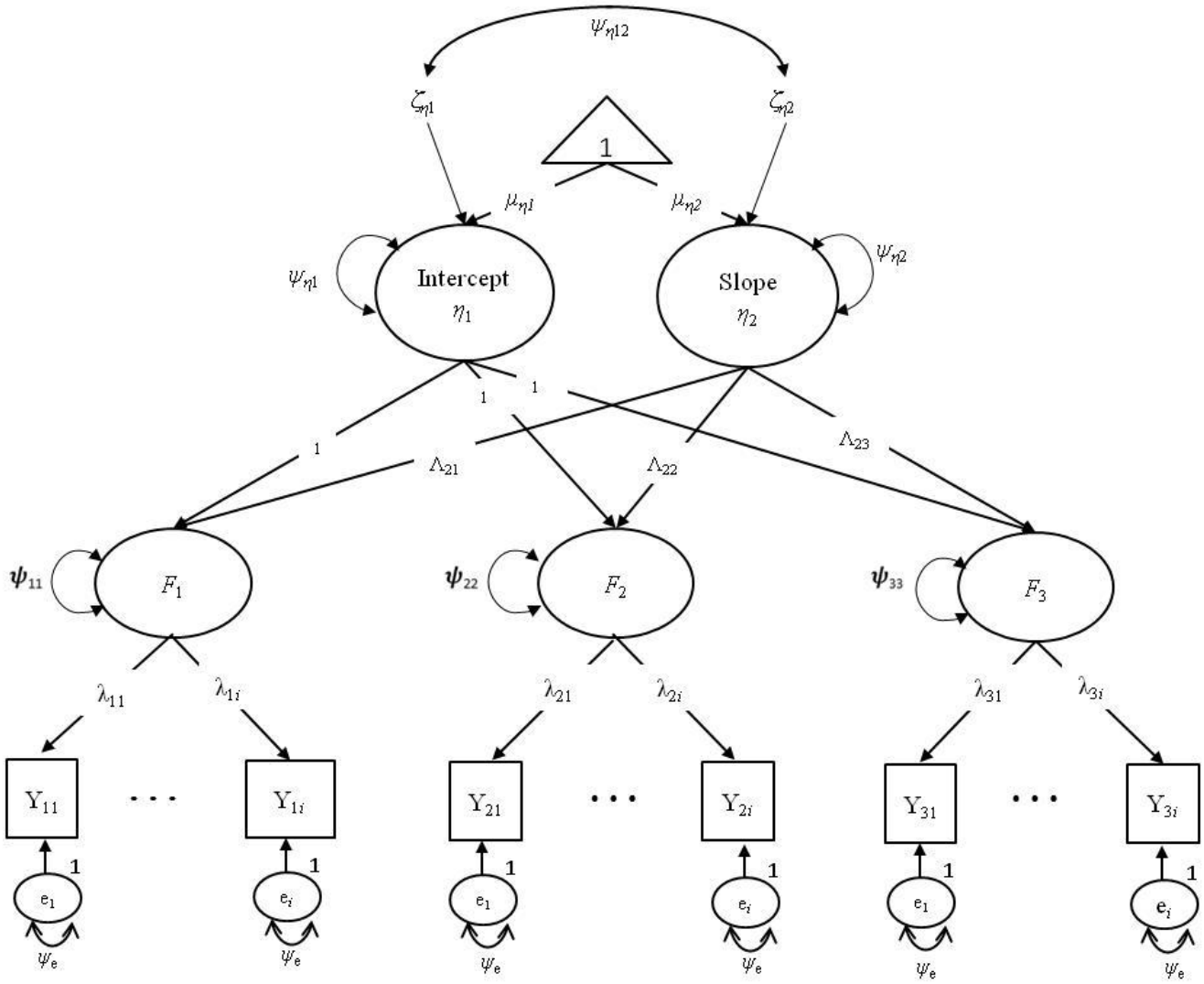

Figure 14. The second-order latent growth models path diagram.

factor loading of F_{tj} at time t; ε_{tj} is measurement errors; η_1 and η_2 are the intercept and slope, that is, second-order factors, respectively; ψ_{tj} are the variance of latent score F_{tj}; $\mu_{\eta 1}$ and $\mu_{\eta 2}$ are the mean of intercept and slope, respectively; and $\zeta_{\eta 1}$ and $\zeta_{\eta 2}$ are the residuals of intercept and slope, respectively. By substituting Equation 103 through 105 into Equation 102, this equation can be rewritten as

$$Y_{tj} = \tau_{ti} + \lambda_{ti}\mu_{\eta_1} + \lambda_{ti}\Lambda_{2t}\mu_{\eta_2} + \lambda_{ti}\zeta_{\eta 1} + \lambda_{ti}\Lambda_{2t}\zeta_{\eta 2} + \lambda_{ti}\psi_{tj} + \varepsilon_{tj}\,. \tag{106}$$

Compared with the LGM model in Figure 13, change in the second-order LGM is modeled in the repeated first-order factors rather than in the manifest variables.

The advantages of this model include that a measurement error of the LGM related to each indicator is confounded with occasion-specific variance, whereas the second-order LGM provides separate estimates for a variance. Furthermore, this model provides more information related to the characteristics of the individual observed variables because of directly modeling the measurement structure of the indicators. Moreover, most importantly, this model allows for testing measurement invariance of the composites over time. In addition, this model contains some potential features. For example, this model should have high reliability, because the occasion-specific variance will be reduced. This is because variance at each time point is separated into residual variance and factor variance (i.e., true score variance) so that it leads to greater statistical power (Newsom, 2015).

The LGM model, in fact, can be extended to other elements of structural models, such as a nonlinear growth model (Newsom, 2015) or an autoregressive latent trajectory model (Bollen & Curran, 2004, 2006). McArdle (1988) has suggested that the LGM model can potentially be integrated with IRT in order to obtain a rigorous basis for meaningful scaling. In this way, integrating with IRT models would provide advantages over traditional approaches.

These advantages include examining measurement invariance, having item and person statistics on the same scale, and using items to discriminate among respondents based on latent abilities (de la Torre & Patz, 2005; Embretson & Reise, 2000; Hsieh et al., 2013).

A number of studies provide empirical results and simulation studies using IRT in the context of a longitudinal growth modeling framework (McArdle et al., 2009; Wang et al., 2016; Wilson, Zheng, & McGuire, 2012). McArdle and colleagues (2009) proposed LGM to systematically model linear and nonlinear growth of ability. Wilson, Zheng, and McGuire (2012) describe a latent growth item response model. The Wilson et al. model permits both linear and nonlinear modeling of change in ability. These two models can be viewed as multidimensional IRT models. The theory underlying LIRT and LGM approach has been available for some time; however, to date, integrating longitudinal IRT (LIRT) models and LGM is a new model.

The LIRT-LGM explores growth in a latent variable of interest based on multiple repeatedly measured observed (manifest) indicators at each occasion using IRT models. The LGM, as used in this study, is based on the second-order LGM. The LIRT model is modified based on McArdle and Nesselroade's (1994) latent change score. As stated above, latent change scores include different models, such as a dual change score model (DCSM). In this study, it is a modified linear change score model (LCSM), which is equivalent to a linear latent growth curve model (McArdle & Nesselroade, 2014). A general assumption for the LIRT-LGM model is that this model assumes each observed variable U_{ij} follows a binomial distribution, and the latent continuous variables (θ_{jt}) underlying the binary outcomes on the item level are assumed to follow a normal distribution. In addition, the measurement errors and residuals in this model are assumed to be zero, and the thresholds and loadings are assumed to be equal over time.

For the LIRT-LGM model, U_{ijt} (i = 1, …, n) is the nth observed scores for respondent j at time t; τ_{it} is the intercept of item i at time t, normally τ_{ti} set to 0 and not shown on Figure 15; η_1 and η_2 are the intercept and slope, that is, second-order factors, respectively; $\Delta\theta_{jt}$ is a latent change score, where $\Delta\theta_{jt} = \eta_2$; Λ_{2t} is a matrix of loadings of the slope (i.e., second-order factors), reflecting the hypothesized growth pattern of the latent continuous variable θ_{jt}; θ_{jt}, is composed of first-order factors and is a latent continuous variable for respondent j at time t; λ_{ti} is the loading of first-order factors of item i at time t; $\mu_{\eta 1}$ and $\mu_{\eta 2}$ are the means of intercept and slope of the growth parameters, respectively; and $\zeta_{\eta 1}$ and $\zeta_{\eta 2}$ are the residuals of intercept and slope, respectively. $\psi_{\eta 1}$ and $\psi_{\eta 2}$ are the variances of intercept and slope, respectively; $\psi_{\eta 12}$ is the covariance of the intercept and slope; and ψ_{11} to ψ_{33} are the variances of latent continuous variables θ_{jt}. This study is interested in the mean slope, which indicates the average amount of change per unit of measurement occasion, and the variance slope. We assume here that indicators of the θ_{jt} are measured on three equal-interval occasions. The equation of Figure 15 can be given as

$$U_{ijt} = \tau_{ti} + \lambda_{ti}\theta_{jt}, \tag{107}$$

where

$$\theta_{jt} = \eta_1 + \Lambda_{2t}\eta_2 + \psi_{jt}, \text{ where } \eta_2 = \Delta\theta_{jt}, \tag{108}$$

$$\eta_1 = \mu_{\eta 1} + \zeta_{\eta 1}, \tag{109}$$

$$\eta_2 = \mu_{\eta 2} + \zeta_{\eta 2}. \tag{110}$$

By substituting Equation 108 through 110 into Equation 107, Equation 107 can be rewritten as

$$\begin{aligned} U_{ijt} &= \tau_{ti} + \lambda_{ti}\mu_{\eta_1} + \lambda_{ti}\zeta_{\eta 1} + \lambda_{ti}\Lambda_{2t}\mu_{\eta_2} + \lambda_{ti}\Lambda_{2t}\zeta_{\eta 2} + \lambda_{ti}\psi_{tj}, \\ &= \tau_{ti} + \lambda_{ti}\mu_{\eta_1} + \lambda_{ti}\Lambda_{2t}\mu_{\eta_2} + \lambda_{ti}\zeta_{\eta 1} + \lambda_{ti}\Lambda_{2t}\zeta_{\eta 2} + +\lambda_{ti}\psi_{tj}. \end{aligned} \tag{111}$$

One of the assumptions of this model is that the item parameters (i.e., difficulty and discrimination parameters) of the factor indicator are equal over time; thus, the subscript *t* will be eliminated, and Equation 111 will be reduced to

$$U_{ijt} = \tau_i + \lambda_i \mu_{\eta_1} + \lambda_i \Lambda_{2t} \mu_{\eta_2} + \lambda_i \zeta_{\eta 1} + \lambda_i \Lambda_{2t} \zeta_{\eta 2} + + \lambda_i \psi_{tj} \,. \qquad (112)$$

The concept of the item parameters remaining equal over time (i.e., measurement invariance) is the usual IRT assumption and is important for investigating change across time points in an underlying latent construct. Measurement invariance will be discussed in the next section.

2.5.1 Measurement Invariance

In a longitudinal study, additional hypotheses should be tested because the same observed variables are measured at repeated time points, but the same constructs are not assured to be measured at each time point. Thus, changes in measured variables can only be interpreted if item parameters are not different across time points. This condition is referred to as measurement invariance. It is an assumption of IRT and means that IRT item parameter estimates do not change when students from the same groups are measured across time on the same items (Cohen, Bottge, & Wells, 2001). For example, item difficulty (b_i) is assumed to be equal over measurement occasions (i.e., $b_{i1} = b_{i2} = \ldots = b_{it}$). Measurement invariance, in this sense, means that the values of parameters are on the same scale over measurement occasions (Ferrer et al., 2008; Meredith, 1993).

Meredith (1993) differentiated between nonmetric (or configural) and metric measurement invariance. Nonmetric measurement invariance indicates that the factor loadings are equal at each time point. Metric measurement invariance can be classified into weak,

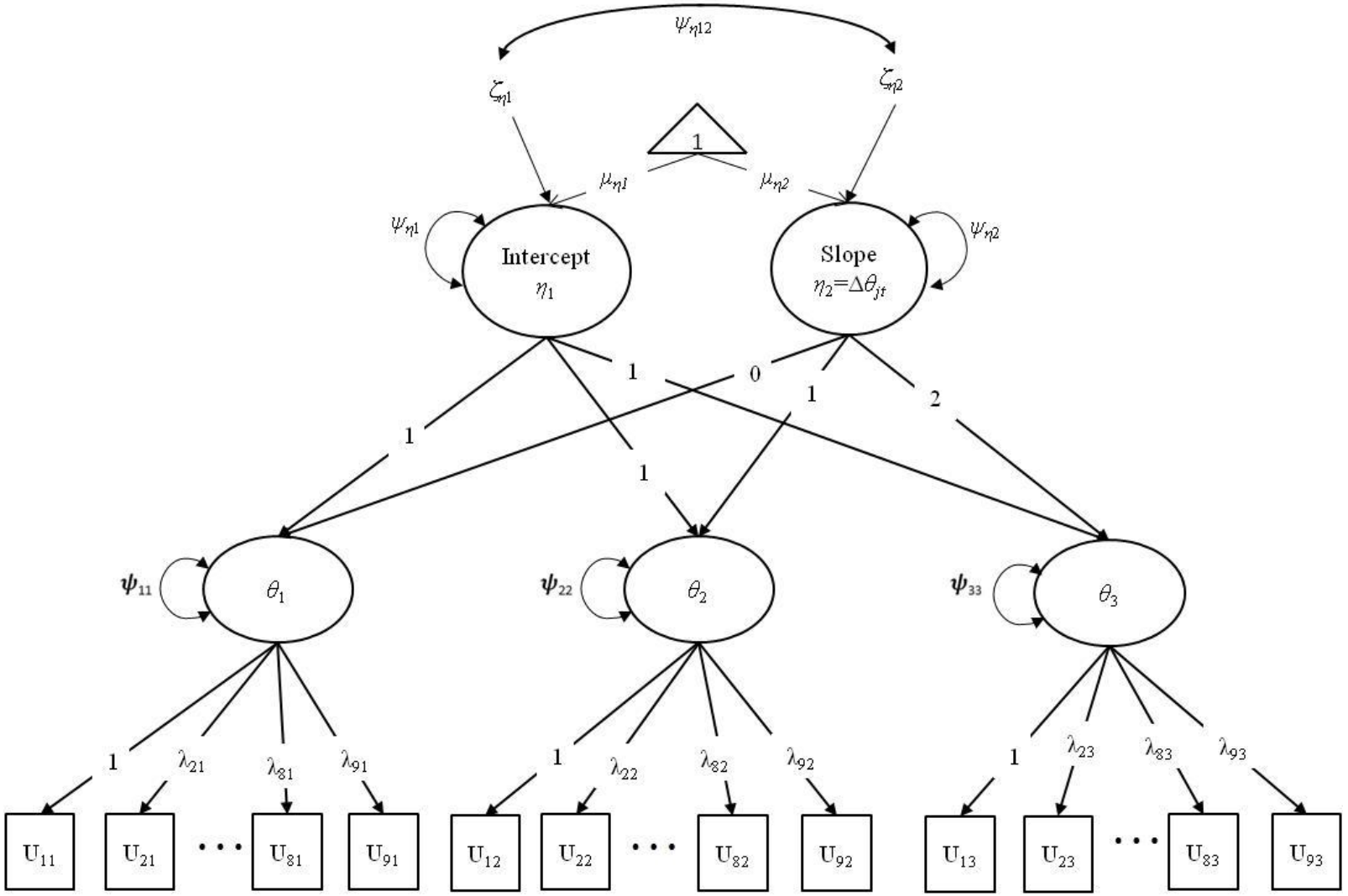

Figure 15. The LIRT-LGM path diagram.

strong, and strict measurement invariance. Weak measurement invariance is the basic assumption of measurement invariance and requires that the factor loading of each indicator must have the same numerical value across time points. Weak invariance is employed in the LIRT-LGM model. That is, factor loadings in Equation 107, for example, are assumed to be equal over the three-time points in the illustration here:

$$\lambda_{11} = \lambda_{12} = \lambda_{13}. \tag{113}$$

Strong measurement invariance requires that the factor loading and measurement intercept of each indicator have the same value over time. Strict measurement invariance requires the factor loading, measurement intercept, and item residual variance of each indicator to be equal over time. However, strict measurement invariance cannot hold while modeling change because heterogeneous variance over time points is often the case. This means that time effects are confounded by time-related increases in the variance (Ferrer et al., 2008; Sayer & Cumsille, 2001).

To examine the assumption of measurement invariance, studies (e.g., Hofer, Thorvaldsson, & Piccinin, 2012) suggest that measurement invariance can be tested by structural equation modeling (SEM). For instance, we estimated a model that freed factor loading as a baseline model (i.e., unconstrained model), and we subsequently estimated a model that constrained the factor loading to be equal. The chi-square difference ($\Delta\chi^2$) test (Bollen, 1989; Cheung & Rensvold, 2002) was used to determine the model fit differences. The $\Delta\chi^2$ can be computed by

$$\Delta\chi^2 = \chi_c^2 - \chi_{uc}^2, \tag{114}$$

$$\Delta df = df_c - df_{uc}, \tag{115}$$

where χ_c^2 indicated the value of constrained model, χ_{uc}^2 denoted the value of unconstrained model, and df_c and df_{un} were the degree of freedom of the constrained model and unconstrained model, respectively. If $\Delta\chi^2$ between the models is not significant, this indicates measurement invariance. In addition to structural equation modeling, measurement invariance can also be tested under IRT by obtaining item parameter using either concurrent or separate calibration (Cho et al., 2013).

If measurement invariance fails to hold, this will cause two problems. First, if factor loadings and intercepts are freely estimated and are found to not be invariant, it is not certain whether the same latent construct can even be measured over time. Hence, the change across time points might not represent quantitative growth in the construct. Rather, it may be reflecting shifts in the nature of the construct over time. Second, choosing different referent indicators for the scaling of the latent factors may change the model fit substantially as well as the growth parameter estimates under partial measurement invariance. Therefore, full measurement invariance is a necessary condition for a valid interpretation of change in latent variables with multiple indicators for a model (Ferrer et al., 2008). If measurement invariance fails to hold, meaning respondents' respond differently to the items over time, the factor means, as a result, cannot be compared reasonably and would be difficult to interpret. When measurement invariance is violated, by choosing other IRT models, researchers may be able to find out which items yielded a poor fit. Researchers could then remove these items from the scale and reevaluate the fit (Millsap, 2010). When the assumption of measurement invariance is met, this means all items are placed on the same scale over time.

2.5.2 The Strengths and Limitations of LIRT-LGM

There are several advantages to using LIRT-LGM for evaluating growth. First, this model allows researchers to simultaneously investigate the invariance, which includes factor structure, factor loadings, and intercepts, and item properties across times. If invariance holds, researchers would be justified in having more confidence that the same latent construct is being measured at each occasion (Ferrer et al., 2008). Second, the model employs multiple indicators rather than a single indicator. This allows researchers to investigate longitudinal change in the latent construct, thus avoiding the limitations and problems implicit with using composite scores. Third, this model can be extended to other models, such as a dual change model or a mixture growth model, if one employs at least four-time points. Fourth, in addition to employing MLE, Bayesian estimation can also be used when sample sizes are small and when responses are perfect or zero response.

In addition to the several advantages, this model also has some limitations. First, referent identification is difficult to choose. The results for the second-order factors could vary when referent identification is problematic. Second, one should have knowledge of the framework of IRT. Third, because of complex mathematic computations, the analysis generally requires more time than simpler growth or IRT models.

2.5.3 The Comparison of the LGM and LIRT-LGM

LGM	LIRT-LGM
LGM estimates change based on a single observed variable where it is summed or average in a set of items to create an index at each time point.	LIRT-LGM estimates change based upon observed multiple indicators with IRT methods at each time point.
LGM is unable to assess measurement invariance.	LIRT-LGM is able to measure measurement invariance.
LGM does not include person- and item- statistics on the same scale.	LIRT-LGM is able to have person- and item-statistics on the same scale.
LGM is unable to use items to discriminate among respondents based on the latent abilities.	LIRT-LGM is able to use items to discriminate among respondents based on the latent abilities.
LGM does not require large sample sizes.	LIRT-LGM requires large sample sizes.

2.6 Research Question and Rationale

Although the theory of LGM and LIRT approaches has been used for a long time, LIRT-LGM is a new model. Thus, additional work to determine the performance of the LIRT-LGM under various conditions is warranted. This study would provide a clear understanding of the performance of the LIRT-LGM under several practical testing conditions including different item lengths, sample sizes, and effect sizes using simulated data so that researchers are able to determine when and how this model may be appropriate to use.

The mean and variance of the slope were of primary interest in this study; thus, the performances of two unconditional models were compared by their mean and variance. Empirical data and simulated data were used in this study. Simulated data were generated with all items present at each measurement occasion. This study attempted to answer the following questions:

1. Do depressive symptoms change in girls' early adolescence? Is this change heterogeneous? This study hypothesized that depression of adolescents would change over time, and it is not heterogeneous in change over time.

2. Are there differences between the performance of LGM and LIRT-LGM?

The major shortcoming of the LGM is that it lacks the ability to model measurement invariance. If the measurement invariance does not hold, the results could be overestimated or underestimated. On the other hand, the LIRT-LGM not only can model measurement invariance but also can have person- and item- statistics on the same scale. Thus, in this study, we hypothesized that the performance between both of these models would differ.

3. Does the 2PL model fit better than the 1PL model?

In this study, we further hypothesized that the 2PL is a fit model because 2PL includes item difficulty and item discrimination. Item discrimination describes that an item can differ between a respondent having an ability level below or above the item location. Thus, this model is useful to refer to the latent trait as the ability common to the *n* items on the test or scale.

4. How do different item lengths and sample sizes influence Type I error using the Monte Carlo simulation?

This study hypothesized that Type I error would be controlled by larger sample sizes and more items.

5. How do different item lengths, sample sizes, and effect sizes influence power?

In this study, we hypothesized that larger sample sizes and more items would have higher power. In addition, when effect sizes increase, power would be high.

CHAPTER 3

METHOD

3.1 Research Structure

This study compared the performance of the LGM and LIRT-LGM models for measuring growth in empirical and simulated data. The empirical study used empirical data to analyze the depressive symptoms of African-American adolescent girls using both the LGM and LIRT-LGM models. The simulation study used a Monte Carlo simulation to generate data where all indicators were present at three-time points in order to compare the overall performance of the LGM and LIRT-LGM models. The simulation study evaluated the performance of two models under three different conditions: test length, sample size (*N*), and effect size (*ES*). Several conditions were manipulated to consider the impact on the Type I error and power. Before examining Type I Error and power, we conducted a recovery study. We placed the estimated and generating parameters on the same scale in order to calculate root mean square error (RMSE), and bias, using the linear equating for the LGM model and using the mean/mean transformation method for the LIRT-LGM model. The analyses of both the empirical study and the simulation study were conducted using M*p*lus 7.4 (Muthén & Muthén, 2011). The research structure was shown in Figure 16.

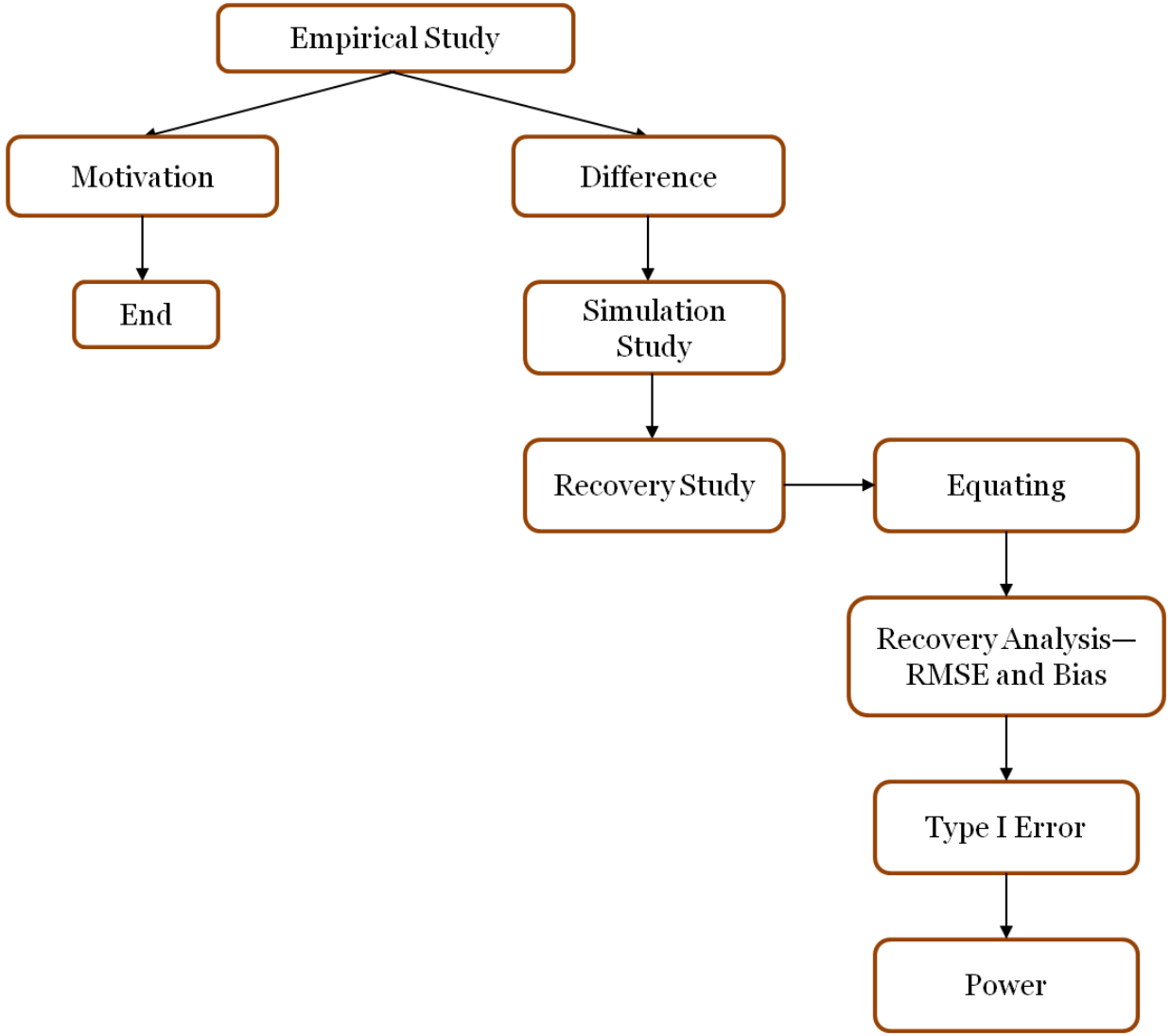

Figure 16. The research structure.

3.2 The Empirical Study

Background

Depression is the leading cause of disability, and one of the top three causes for the burden of diseases in the world (Fried & Nesse, 2015). It has resurged as an important topic in social and psychological development. Indeed, depression not only results in health-related problems, but also increases the risk of suicide (McGirr et al., 2007; O'Kearney et al., 2009).

Traditionally, depression is calculated by adding up scores for different depressive symptoms to create a sum-score. High scores indicate high levels of depression. Previous studies (LeMoult et al., 2015; Ge et al., 2003; Hayward et al., 1997; Keenan et al, 2008; Petersen, Sarigiani, & Kennedy, 1991) have provided evidence that early physical maturation was significantly related to elevate depressive symptoms among girls. Early-maturing girls have consistently been found to be more likely to exhibit depressed moods than their on-time or late-maturing peers. However, Ge et al. (2003) and Jones and Bayley (1950) found that early-maturing boys' depressive symptoms were not necessarily related to age of maturation, mainly because boys enjoyed social advantages over their on-time or late-maturing peers. Thus, depressive symptoms are more likely to be exhibited among adolescent girls. In this study, the focus was on adolescent girls at a time during which individual differences in depression were likely to be most clearly detected.

Repeated measures ANOVA had traditionally been used to analyze depression (Petersen et al., 1991). This approach has some important limitation, however, including inability to deal with missing data. Thus, many studies have employed LGM to analyze depression (Lindhorst & Oxford, 2008; Pettit et al., 2011). This study used both the LGM and LGM-LIRT to analyze depressive symptoms. To better understand how depressive symptom changes actually take

place, this study used the longitudinal data from the Family and Community Health Study (FACHS) that was designed to identify neighborhood and family processes that contribute to the development of African-American children (Simons et al., 2011).

Research Design

Data. This study used three time points of data collected for the Family and Community Health Study (FACHS), a multi-site (i.e., Iowa and Georgia) investigation of neighborhood and family effects on mental health and development of African-American children (Beach et al., 2012; Simons et al., 2011). The sampling strategy was intentionally designed to generate a data set with families representing a range of socioeconomic statuses and a wide variety of neighborhood settings. In both Iowa and Georgia, households were randomly selected from the sampling frame using rosters of fifth-grade students in the public school systems. When a household did not interest in participating in the project, it was removed from the rosters, and other households were randomly selected until the required number of households had been recruited (Simons et al., 2011).

During the data collection procedures, the interviewers received a month of training in the administration of the self-report instruments. When the families' schedules allowed, two home visits were made to each family within seven days. Each visit took on average two hours to complete. Informed consent was obtained during the first visit. The primary caregivers (PCs) agreed to participate along with the children in their care in this interview, and the children also consented to participate in the interview. Self-report questionnaires were administered to the PCs and the children in an interview format during each visit because of literacy concerns. Each interview was conducted privately between one participant and one interviewer without any

other family members present. In the interviews, the instruments were presented on a laptop computer. A series of questions were shown on the screen so both the interviewer and participant could see the questions. The interviewer read each question aloud and entered the responses of the participants on the computer. Identical procedures were used in other time points (see below) for collecting data (Ge et al., 2003).

Sample. The FACHS sample consisted of 889 African-American children (411 boys and 478 girls) with their PCs during the first time point. A majority of the PCs were female (829), and only 60 were men. Eighty-four percent of the mothers in the sample were biological mothers of the targets. Further, 44% of the mothers were identified as single parents. The educational levels of the PCs were varied, ranged from less than high school diploma (19%) to a bachelor's or graduate degree (9%). At the study's inception, about half of the sample resided in Georgia (n = 422) and the other half in Iowa (n = 467). The first time point of data, collected from 1997 to 1998, included interviews with 889 respondents. The second time point of the data, collected from 1999 to 2000, included re-interviews with 779 respondents. The third time point of the data, collected from 2001 to 2002, included interviews with 767 respondents (Simons et al., 2011).

The sample in this example was drawn from the host study sample and had no missing data for the variables of interest. Thus, this sample consisted of 381 African-American girls between ages 10 and 15 years.

Measures

Depressive symptoms. Depressive symptoms were evaluated with the Diagnostic Interview Schedule (Robins, Helzer, Croughan, & Ratcliff, 1981). This measure focused on the list of symptoms of depression in the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) for major or minor depressive episodes. We assessed all nine depressive symptoms based on criteria in the DSM-IV manual. The items rated on a dichotomous scale using diagnostic algorithms that correspond to the DSM-IV criteria were developed by the Division of Child and Adolescent Psychiatry at Columbia University (Shaffer et al., 1993). Responses were coded "present" as 1 and "absent" as 0 for each of the items describing one of the nine depressive symptoms. The depressive symptom score (Belden et al., 2015; Fried & Nesse, 2015) was computed by the sum of the number of "present" responses for the nine diagnostic symptoms. Thus, a score of at least 5 of these 9 corresponds to a respondent's answers for diagnosing a respondent who has depression. Note that, one of the depressive symptoms must be depressed mood. The reliability of the scale was similar across the three time points, with a coefficient's alpha of r_{α} = .86 at Time 1, r_{α} = .89 at Time 2, and r_{α} = .88 at Time 3. A list of the nine items measuring depressive symptoms is presented in Appendix A.

Method of Analysis

Hypotheses were tested for the LGM and LIRT-LGM models with the M*p*lus 7.4 program (Muthén & Muthén, 2011) to measure the depression of African-American adolescent girls across three-time points. Both models used an unconditional model analysis measuring a growth curve for three-time points, calculated the intercept and slope of average growth curves and compared the performance of these two models. The data were described by the LGM and LIRT-

LGM, having a mean and variance parameter. For the LGM, this study estimated two latent variables, intercepts and slopes, using the indicators from each of the three time points under the unconditional model. The first latent variable is an intercept, which is constant for any given individual across time, and represents the estimated level at Time 1. It is denoted in the both LGM and LIRT-LGM model as intercept η_1. The loadings of the intercept are fixed at 1 at all time points. The second latent variable is a slope, which indicates an individual's change. It is denoted in the both LGM and LIRT-LGM model as slope η_2. The loading of the slope is fixed at 0 at Time 1, at 1 at Time 2, and at 2 at Time 3.

The three observed variables U_1 to U_3 (see Figure 13) were constructed as the sum of the nine items in order to create an index of depressive symptoms for use with the LGM model. This model treats depressive symptoms as an observed variable. For the LIRT-LGM model, the first-order factors (θ_1 through θ_3) incorporated the multiple indicators into the model. Measures of depressive symptoms were treated as categorical variables with thresholds constrained to equality across time points in the LIRT-LGM. Maximum likelihood was used to estimate the unknown and corresponding parameters. The results of these models were used to describe the mean trajectory for depression.

3.3 The Simulation Study

The Monte Carlo simulation study was used to evaluate the performance of the LIRT-LGM model compared with the LGM model. Specifically, the differences in test length, effect size, and sample size were investigated. In this investigation of the LIRT-LGM model, a linear growth pattern was simulated with complete data. Curvilinear trajectory conditions or missing data conditions were not considered. In this simulation study, the Type I error and power of the

LIRT-LGM were compared to those of the LGM. In addition, prior to determining the Type I error, a recovery analysis was first conducted. The measures for the recovery analysis included correlation coefficient (*r*), root mean square error (RMSE) and bias. For those conditions in which the Type I Error was controlled, power is evaluated.

The Simulation Design

For LGM and LIRT-LGM, various conditions were manipulated in the comparisons of the LGM and the LIRT-LGM models. This study simulated two different test lengths (10 and 30 items), three sample sizes (100, 500, and 1,000), and four effect sizes (.0, .1, .2, and .3) for each model. One thousand replications were simulated for each condition consistent with Leite (2007). Thus, there were 24 ($2 \times 3 \times 4$) independent simulation conditions and 1,000 replications for each condition resulting in a total of 24,000 datasets being generated.

Data Generation and Analysis

The M*p*lus 7.4 program (Muthén & Muthén, 2011) was used for all statistical simulations and analyses. The M*p*lus syntax to generate longitudinal data and to analyze each model is provided in Appendices F to K.

When generating data, the seed option was used in M*p*lus to start the random draw for each experimental condition. Item responses were generated based on the binomial distribution, and dependent variables were simulated to be categorical.

The generating means of the intercepts were fixed at zero, consistent with Muthén and Muthén (2011). Also, the generating mean of the slope was set to zero, since it was used to inspect the Type I error. The generating intercept-slope covariance was fixed at zero consistent

with Hertzog, von Oertzen, Ghisletta, and Linfenberger (2008). The generating variances of intercept and slope followed Muthén and Muthén's (2011) example, and they were set to 1 and .2, respectively. Thus, the covariance matrix was

$$\psi = \begin{bmatrix} 1 & 0 \\ 0 & .2 \end{bmatrix}. \tag{116}$$

For three latent variables (θ_1 through θ_3), the generating mean and variance were estimated fixed at value of 0 and 1. This standardizes the scale of the variable to a normal distribution with mean of zero and standard deviation of 1 for those latent variables. The values of the loadings and thresholds were generated based on Embretson and Reise (2000). Using a unidimensional model, the population values of thresholds were selected with increments of 0. 5, that is [-2.0, -1.5, -1.0, -0.5, 0, 0, 0.5, 1.0, 1.5, 2.0], and population loadings increase in 0.10 increments, that is [1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9]. Note that this study fixed the first loading to 1 in order to give the latent variables an interpretable scale, otherwise the scale of the latent variables could not be estimated (Hox & Bechger, 1998). The values of the parameters of the thresholds and loadings of the same items were fixed over repeated measures. The significant level in terms of the Type I error rates and power were set to .05. Furthermore, the value of the loadings of the slope were fixed starting from zero on the first measurement occasion, with increases of one (i.e., 0, 1, and 2), and the loadings of the initial level were all fixed at one. Conditions were simulated with weak measurement invariance based on the invariance taxonomy proposed by Meredith (1993). In a weak measurement condition, this study constrained the loadings and thresholds to be the same across three occasions to identify the invariant to be held with the data (Meredith, 1993). The parameters estimates and standard error

were estimated using maximum likelihood. Note that this study only focused on completed case analysis; thus, it did not consider the effects caused by missing data.

Analysis of the LGM used the same technique as an empirical study; that is, each observed variable at each time point is constructed as the sum of a set of items to create an index for that time of measurement. As stated above, this study was interested in the mean and variance of the slope of the second-order factors. Effect sizes were set at .0 for detecting Type I errors and set to .1, .2, and .3 for detecting power, and the variance of the slope was set to .2 following Muthén and Muthén (2011). For the analysis of the LIRT-LGM model, the mean of the intercept was fixed at zero and the variance of the intercept was set to 1 as suggested by Muthén and Muthén (2011). The mean and variance of the slope were set at the same values as the LGM for detecting the Type I error and power. The mean and variance of the first-order factors (i.e., θ_1 through θ_3) were fixed to 0 and 1, respectively.

Recovery Study

Correlation coefficient (r), root mean square error (RMSE), and bias were used for analysis of model recovery. r is an index of the degree of linear relationship between the generating and estimated parameter, ranging from -1 to 1. There is no relationship existed when $r = 0$; however, when $r = 1$ or -1, there is existed a perfect positive or negative linear relationship. r was calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (117)$$

where x_i and y_i were the estimated parameter and the generating parameter for item i, respectively; and $\bar{x}$ and $\bar{y}$ are the means of the estimated parameter and the generating parameter for item i, respectively.

The root mean square error (RMSE) and bias were used to assess the recovery of item parameters between the generating parameter and the estimating parameter. The smaller the RMSE values, the better the estimation accuracy. Similarly, a zero value of bias indicated unbiased parameter estimates. We evaluate RMSE and bias for difficulty parameters as below:

$$\text{RMSE}_{(bi)} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{r=1}^{R}(b_i - \hat{b}_{ir})^2}{Rn}}, \tag{118}$$

$$\text{Bias}_{(bi)} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{R}(b_i - \hat{b}_{ir})}{Rn}, \tag{119}$$

where b_i is the generating item difficulty parameter for item i, and $\hat{b}_{ir}$ is the estimated item difficulty parameter for item i and replication r, where $r = 1, \ldots, R$; and n is the number of items where $i = 1, \ldots, n$. RMSE and bias for discrimination parameters were calculated as below:

$$\text{RMSE}_{(ai)} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{r=1}^{R}(a_i - \hat{a}_{ir})^2}{Rn}}, \tag{120}$$

$$\text{Bias}_{(ai)} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{R}(a_i - \hat{a}_{ir})^2}{Rn}, \tag{121}$$

where a_i is the generating item discrimination parameters, and $\hat{a}_{ir}$ is the estimated item discrimination parameter for item i and replication r. However, before calculating RMSE and

bias, the generating and estimated parameters were placed on the same scale. This dissertation used the linear equating for the LGM model as below (Macro, 1977; Kolen & Brennan, 2004)

$$\frac{E-\mu_E}{\sigma_E}=\frac{G-\mu_G}{\sigma_G}, \tag{122}$$

where E was the score on the estimated parameter and G was the score on the generating parameter; $\mu(E)$ and $\mu(G)$ were the means of estimated parameter and generating parameter, respectively; and $\sigma(E)$ and $\sigma(G)$ were the standard deviations of estimated parameter and generating parameter, respectively. Solving for the generating parameter score G will give a formula for adjusting the raw score E on the estimated parameter as below

$$G=L_G\left(E\right)=\left(\frac{\sigma_G}{\sigma_E}\right)E+\left[\mu_G-\left(\frac{\sigma_G}{\sigma_E}\right)\mu_E\right], \tag{123}$$

where $L_G(E)$ indicated the adjusted scores on the estimated parameter would have the same μ and σ as the raw scores on the generating parameter. If $\left(\frac{\sigma_G}{\sigma_E}\right)=$ slope (a) and $\mu_G-\left(\frac{\sigma_G}{\sigma_E}\right)\mu_E=$ intercept (b), Equation 123 can be written as

$$G=aE+b. \tag{124}$$

When a and b are determined, scores for estimated parameter will put on the same scale as scores for generating parameter. We used mean/mean transformation methods for the LIRT-LGM using IRTEQ computer software (Han, 2009) as below (Macro, 1977; Kolen & Brennan, 2004)

$$A=\frac{\mu\left(a_E\right)}{\mu\left(a_G\right)}, \tag{125}$$

$$B=\mu\left(b_G\right)-A\mu\left(b_E\right), \tag{126}$$

where $\mu(a_E)$ and $\mu(a_G)$ were the mean of item discrimination of the estimated parameters and generating parameter, respectively, and $\mu(b_E)$ and $\mu(b_G)$ were the mean of item difficulty of the estimated parameters and generating parameter, respectively. After calculating the values of *A* and *B*, we placed the estimated parameters on the same scale as the generating parameters using the following transformation (Macro, 1977):

$$b_E = \frac{b_G - B}{A}, \tag{127}$$

$$a_E = A(a_G), \tag{128}$$

where b_E and b_G is the item difficulty of the estimated and generating parameters, respectively, and a_E and a_G is the item discrimination of the estimated and generating parameters, respectively.

Test Statistics

Type I Error. The null hypothesis can be rejected based on a level of significance α. The rejection region (RR) indicates the values of test statistics in order to reject the H_0 in favor of the H_a. Thus, when the value of the test statistic falls in the RR, H_0 will be rejected, and H_a will be accepted. On the other hand, when the computed value of the test statistic does not fall in the RR, H_0 will be accepted (Wackerly, Mendenhall, & Scheffer, 2008). A Type I error is defined as the probability of rejecting the true H_0. The Type I error rate should be close to .05 when α = .05. This study was interested in estimating the means of the slope; thus, the null hypothesis was defined as

$$H_o : \mu_{\eta 2} = 0, \tag{129}$$

$$H_a : \mu_{\eta 2} \neq 0 \tag{130}$$

under all conditions (i.e., different test lengths and sample sizes). $\mu_{\eta 2} = 0$ means no growth over repeated measurements. Since a 5% level of significance is commonly used in behavioral and social sciences fields, this study reported Type I error rates relative to that value. The mean of the slope would be expected to be incorrectly significant in 50 of the 1000 samples when $\mu_{\eta 2} = 0$ at a 5% level of significance. Type I error will be a good control when the Type I error rate is equal to the nominal Type I error rate $\alpha = .05$ (Pearson, 1927). The range for the Type I error rate were evaluated using Bradley's (1978) criteria. Bradley indicated two criteria for the range null hypothesis, including stringent and liberal criteria. The stringent criterion stated that the range null hypothesis should lie in $\alpha \pm 0.1\alpha$, and the range null hypothesis of the liberal criterion should lie in $\alpha \pm 0.5\alpha$. This study employed the stringent criterion (i.e., .045 to .055) to study the Type I error rate. When Type I error were controlled by the LGM and LIRT-LGM models, we can examine power.

Power. Rejecting a false H_0 is associated with power. Statistical power is a useful concept for judging the performance of a test. It is given as one minus the probability of making a Type II error, β, (i.e., power = 1- β) (Murphy & Myors, 1998).

Model Selection for LIRT-LGM

−2 Loglikelihood was used to inform the model selection of the 1PL and 2PL for the LIRT-LGM model. This value is distributed as a chi-square with degrees of freedom being the difference between the number of parameters estimated by each model. Significance of the chi-square was evaluated at the .05 level.

CHAPTER 4

RESULTS

In general, the results of the empirical study and simulation study were consistent in that they both showed that the performance of the LIRT-LGM was better than the LGM model. The results of the empirical study will be discussed in the next section.

4.1 Results of the Empirical Study

Of the initial 478 African-American girl adolescents, three did not respond at Time 1, 32 did not respond at Time 2, 29 did not respond at Time 3, two did not respond at Time 1 and Time 3, 30 did not respond at Time 2 and Time 3, and one did not respond during Time 1, Time 2, and Time 3. Thus, the analyses of the empirical study were based on a sample size of 381 African-American adolescent girls with their primary caregivers. The total symptom counts in this sample had a mean of 0.30 (*SD* = 1.09) at Time 1, a mean of 0.40 (*SD* = 1.31) at Time 2, and a mean of 0.49 (*SD*= 1.45) at Time 3. Thus, the mean depressive symptoms scores consistently increased across three-time points as shown in Figure 17.

The DSM-IV specifies nine depressive symptoms, such as depressed mood or irritable and suicidality, and the frequency analyses of each time point were showed in Table 1. Table 1 reported that except for Item 2 and Item 9, depressive symptoms gradually increased over three-time points. The highest frequency of depressive symptoms was Item 1 (depressed mood or irritable; frequency = 4.72, 8.14, and 10.76, respectively), and the lowest frequency of depressive

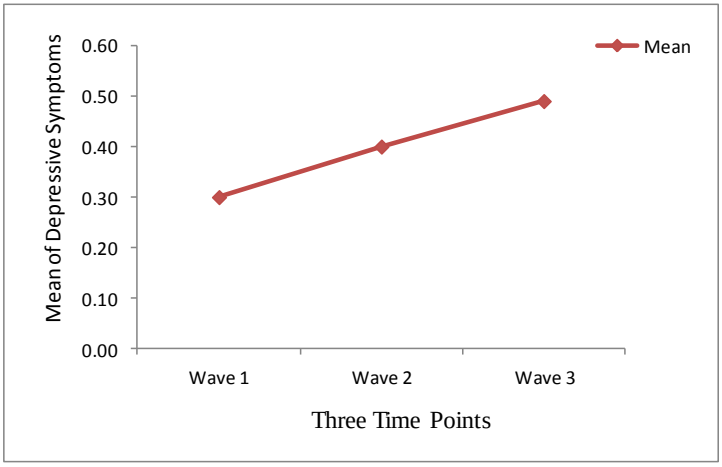

Figure 17. The mean of the depressive symptoms of the empirical study.

symptoms was Item 5 (change in activity; frequency = 1.57, 2.36, and 2.89, respectively) across the three-time points.

4.1.1 Results for the LGM

The unconditional LGM fit the adolescent African-American girls' data well: $\chi^2 = 0.004$, df=1, p=.9528; χ^2 was not significant. The comparative fit index (CFI) was 1.000, indicating a good fit. The root mean square error of approximation (RMSEA) was 0.000, which indicates a good fit, because the acceptable value was smaller or equal to 0.05, and the standardized root mean square residual (SRMR) was 0.001. This is considered indicating a good fit, because the value was less than 0.008. The results of the mean and variance of the intercept and of the slope of the LGM and LIRT-LGM are reported in Table 2.

The negative covariance of the intercept and slope was significant and indicated that African-American adolescent girls' levels of depression were less likely to change across three-time points. The mean of the slope was positive and significantly different from zero, indicating

that the mean of the slope increased between Time 1 and Time 3 at an average rate of .093 points each time point. The estimated mean for the intercept indicated that adolescent girls reported their levels of depression at the first time point. The variances in both intercept and slope were statistically significant, indicating robust individual differences in the trajectories of depression.

4.1.2 Result for the LIRT-LGM

In this model, all loadings and thresholds for the first-order factors (i.e., θ_1 to θ_3) were constrained to be equal across three-time points in order to evaluate measurement invariance. The first-order factors for each factor had a zero mean and a unit variance. The key interest in this model was the mean and variance of the slope. Results for the LIRT- LGM, however, differed from the LGM. For the 1PL and 2PL, the negative covariance of the intercept and slope was not significant, indicating no relationship between the intercept and slope. In addition, the mean of the slope was positive and not significant, reflecting the mean of the slope as having no change from Time 1 through Time 3. Moreover, the variances of the intercept of both the 1PL and 2PL were not significant. The variance of the slope was not significant in the 1PL model, indicating no individual differences in trajectories of depression, whereas the variance of the slope was statistically significant in the 2PL model, indicating individual differences in the trajectories of depression. Loadings and thresholds for the depressive symptoms are listed in Table 3. Item 6, fatigue or lossof energy, had a higher loading (3.423) and the threshold (17.703) for the 2PL, and the 1PL indicated that Item 5 had a higher threshold, indicated easier items to respond.

Table 1

Descriptive Statistics of Depressive Symptoms across the Three Time Points

	Time 1				Time 2				Time 3			
	Presence =1		Absent =0		Present = 1		Absent = 0		Present = 1		Absent = 0	
Symptoms	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
1. Depressed mood or irritable	18	**4.72**	363	95.28	31	**8.14**	350	91.89	41	**10.76**	340	89.24
2. Decreased interest or pleasure	25	6.56	356	93.44	20	5.25	361	94.75	26	6.82	355	93.18
3. Significant weight change (5%) or change in appetite	18	**4.72**	363	95.28	23	**6.04**	358	93.96	27	**7.09**	354	92.91
4. Change in sleep	9	**2.36**	372	97.64	15	**3.94**	366	96.06	18	**4.72**	363	95.28
5. Change in activity	6	**1.57**	375	98.43	9	**2.36**	372	97.64	11	**2.89**	370	97.11
6. Fatigue or loss of energy	9	**2.36**	372	97.64	14	**3.67**	367	96.33	15	**3.94**	366	96.06
7. Guilt/worthlessness	6	**1.57**	375	98.43	9	**2.36**	372	97.64	13	**3.41**	368	96.59
8. Concentration	13	**3.41**	368	96.59	14	**3.67**	367	96.33	19	**4.99**	362	95.01
9. Suicidality	10	2.62	371	97.38	16	4.20	365	95.80	15	3.94	366	96.06

Table 2

The summary of Mean, Variance, and Covariance of the Latent Growth Modeling and Longitudinal Item Response Theory-Latent Growth Modeling

	LGM	LIRT-LGM	
		1PL	2PL
Mean			
Intercept	0.300**	N/A	N/A
Slope	0.093*	0.279	0.310
Variance			
Intercept	0.580**	5.682	6.180
Slope	0.357**	1.577	1.652*
Covariance of the intercept and slope	-0.239*	-1.595	-1.689

Note: *p < .05; **p<.001

Table 3

The Loadings and Thresholds for Depressive Symptoms

	Loading		Threshold	
Item	1PL	2PL	1PL	2PL
1. Depressed mood or irritable	N/A	1.000**	4.652**	4.793**
2. Decreased interest or pleasure	N/A	1.304**	8.337**	6.351**
3. Significant weight change (5%) or change in appetite	N/A	1.922**	8.440**	9.033**
4. Chang in sleep	N/A	2.003**	9.470**	10.48**
5. Change in activity	N/A	2.968**	**10.335****	16.341**
6. Fatigue or loss of energy	N/A	**3.423****	9.661**	**17.703****
7. Guilt/worthlessness	N/A	2.572**	10.209**	14.139**
8. Concentration	N/A	2.092**	9.290**	10.713**
9. Suicidality	N/A	1.310**	9.516**	7.303**

Note: *p < .05; **p<.001

In sum, both the LGM and 2PL of the LIRT-LGM were similar on only one condition: the LGM and the 2PL of the LIRT-LGM models indicated the variance of the slope to be statistically significant. Both models, however, also showed different results on the mean of the slope, the variance of the intercept, and the covariance of the intercept and slope. Hence, the results for the both LGM and the LIRT-LGM models were different for the empirical data. Since most conditions were different, a simulation study was used to investigate the performance of the two models under practical testing conditions.

4.2 Results of the Simulation Study

This study first examined the recovery of the item parameters, means, and variances. Next, the simulation study examined Type I error rates for each of the simulation conditions. Power was then determined for those conditions in which Type I error control was demonstrated.

4.2.1 Recovery Analysis

A recovery analysis was done in order to verify whether M*p*lus accurately recovered the generating parameters under the condition simulated. Correlation, RMSE, and bias were estimated to compare estimates with generating values to assess the recovery of item parameters for the LGM and LIRT-LGM models. In order to compare estimated parameters with the generating parameters using the RMSE and bias, the estimated parameters of both the LGM and LIRT-LGM models were placed on the same scale of the generating parameters using linear equating for the LGM model (Macro, 1977; Kolen & Brennan, 2004) and mean/mean transformation methods for the LIRT-LGM model (Kolen & Brennan, 2004) . Correlations do not require both item parameter estimates to be on the same scale. Mean RMSE and mean bias

for each condition of the LGM and the LIRT-LGM were separately tabulated. The generating and estimated parameters of different test lengths for each condition were tabulated for thresholds and loadings separately. These are presented in Appendices B through E.

The correlations of both the LGM and LIRT-LGM models are presented in Table 4. Correlations for the LGM model ranged from -.812 to .995. Correlations for the 1PL ranged from .994 to .995 and for the 2PL ranged from .999 to 1.000. Correlations for the 2PL version of the LIRT-LGM indicated a perfect positive linear relationship.

The results of the recovery analysis for mean RMSE and mean bias for each condition for thresholds and loadings are shown in Tables 5 through 8. The values of mean RMSE for the LGM model ranged from 0.090 to 0.411. For the LGM model, the mean RMSE values increased, when test length and sample size increased. The mean RMSE values of the LGM were large, in particular, for the 30-item tests with different sample sizes. This indicated poor recovery because of a negative correlation. For the LIRT-LGM model, the mean RMSE values of the thresholds for the 1PL were 0.003 and for the 2PL were close to 0.000. The mean RMSE values of the loadings for the 2PL ranged from closed to 0.000 to 0.001 (see Table 6). The mean RMSE values of thresholds and loadings for the LIRT-LGM model were small, indicating good recovery for the LIRT-LGM model. The results indicated that mean RMSE values of the LGM model were larger than those for the LIRT-LGM model.

Mean bias values for both the LGM and LIRT-LGM models are displayed in Tables 7 and 8. The mean bias values for the LGM model ranged from 0.003 to 0.013 and for the LIRT-LGM model were close to 0.000. The small bias values indicated accurate parameter estimates. Thus, these values indicate good recovery of the generating parameters for the LIRT-LGM

model across all simulation conditions. The results showed that the mean bias values of the LGM model were larger than the LIRT-LGM.

Table 4

The Correlation for the LGM and LIRT-LGM Models

		LGM	LIRT-LGM		
			1PL	2PL	
Item	*N*		*b*	*a*	*b*
10	100	-.990	.994**	.999**	.999**
	500	.904	.994**	.999**	.999**
	1000	.995	.994**	.999**	1.000**
30	100	-.812	.995**	.999**	.999**
	500	.919	.995**	.999**	.999**
	1000	.926	.995**	.999**	.999**

Note: *p<.05; **p<.001; *b* = Threshold (or Item Difficulty); *a* = Loading (or Item Discrimination).

Table 5

The Mean RMSE for the LGM Model over 1000 Replications

		LGM
Item	*N*	RMSE
10	100	0.090
	500	0.090
	1000	0.091
30	100	0.409
	500	0.410
	1000	0.411

Table 6

The Mean RMSE for the LIRT-LGM Model over 1000 Replications

		LIRT-LGM		
		1PL	2PL	
		RMSE	RMSE	
Item	*N*	*b*	*a*	*b*
10	100	0.003	0.001	0.000
	500	0.003	0.000	0.000
	1000	0.003	0.000	0.000
30	100	0.003	0.000	0.000
	500	0.003	0.000	0.000
	1000	0.003	0.000	0.000

Note: *b* = Threshold (or Item Difficulty); *a* = Loading (or Item Discrimination).

Table 7

The Mean Bias for the LGM Model over 1000 Replications

		LGM
Item	*N*	Bias
10	100	0.003
	500	0.003
	1000	0.003
30	100	0.013
	500	0.013
	1000	0.013

Table 8

The Mean Bias for LIRT-LGM Models over 1000 Replications

		LIRT-LGM		
		1PL	2PL	
		Bias	Bias	
Item	*N*	*b*	*a*	*b*
10	100	0.000	0.000	0.000
	500	0.000	0.000	0.000
	1000	0.000	0.000	0.000
30	100	0.000	0.000	0.000
	500	0.000	0.000	0.000
	1000	0.000	0.000	0.000

Note: *b* = Threshold (Item Difficulty); *a* = Loading (Item Discrimination).

The mean RMSE and mean bias for the means and variances of the slope for the LGM and LIRT-LGM models are presented for different test lengths, sample sizes, and effect sizes in Tables 9 through 11. Recall that, in the simulation conditions, the generating means of the slope were set to .0, .1, .2, and .3, and the generating variance of the slope were set to .2. For the LGM model, the values in Table 9 showed that the estimated means of the slope ranged from -0.004 to 1.576. Values differed slightly from the generating mean. The mean RMSE values of the mean slope ranged from closed to 0.000 to 0.040. These were small when the generating mean was equal to .0 with different test lengths and sample sizes. However, when the generating mean was equal to .3 for the different test lengths and sample sizes, the mean RMSE values of the mean were larger than other conditions. The values of the mean bias of the mean slope ranged from -0.001 to closed to 0.000. The small mean bias indicated the accurate parameter estimates. The estimated variances of the slope for the LGM models ranged from 0.747 to 7.131. These values

differed from the generating variances. The mean RMSE values of the variance slope ranged from 0.017 to 0.218. The test length of 30 yielded the largest mean RMSE values for variance of the slope. The values of the mean bias for the variance of the slope ranged from 0.001 to 0.007.

For the LIRT-LGM model, Table 10 showed that the estimated variances of the slope for 1PL model ranged from 0.006 to 0.314. Both mean RMSE and mean bias values for the variance slope of the 1PL were close to 0.000, and the small value for both mean RMSE and mean bias indicated good recovery. The estimated variances of the slope for 1PL model ranged from 0.196 to 0.257. The mean RMSE values were around 0.002 only when the sample size was 100. When sample sizes increased, the mean RMSE values were close to 0.000. The mean bias values of variance slope also were close to 0.000. This indicated good recovery. Table 11 showed the estimated variances of slopes for the 2PL model ranged from -0.005 to 0.300. Both mean RMSE and mean bias values for the variance slope of the 2PL were close to 0.000, indicating good recovery. The estimated variances of the slope for the 2PL model ranged from 0.199 to 0.241. The mean RMSE values were around 0.001 when the sample size was 100. When sample sizes increased, the mean RMSE values were close to 0.000. The mean bias values of variance slope were close to 0.000 for all simulation conditions, indicating good recovery.

In sum, the results presented here indicate that the item parameters, including thresholds and loadings, of the 2PL version of the LIRT-LGM model for different test lengths and sample sizes were reasonably well recovered. The means and variances of the LIRT-LGM model were recovered better than for the LGM model. In addition, the mean RMSEs and mean biases for the LIRT-LGM were smaller than for the LGM model, further indicating that the LIRT-LGM was recovered well.

Table 9

The Recovery Analysis for Means and Variances for the LGM Model

		LGM							
Item	*N*	G. Mean	E. Mean	RMSE	Bias	G. Var	E.Var	RMSE	Bias
10	100	0.0	-0.004	0.000	0.000	0.2	0.747	0.017	-0.001
		0.1	0.172	0.002	0.000	0.2	0.754	0.018	-0.001
		0.2	0.348	0.005	0.000	0.2	0.774	0.018	-0.001
		0.3	0.522	0.007	0.000	0.2	0.793	0.019	-0.001
	500	0.0	0.003	0.000	0.000	0.2	0.760	0.018	-0.001
		0.1	0.178	0.002	0.000	0.2	0.768	0.018	-0.001
		0.2	0.353	0.005	0.000	0.2	0.777	0.018	-0.001
		0.3	0.524	0.007	0.000	0.2	0.785	0.018	-0.001
	1000	0.0	0.002	0.000	0.000	0.2	0.756	0.018	-0.001
		0.1	0.178	0.002	0.000	0.2	0.765	0.018	-0.001
		0.2	0.353	0.005	0.000	0.2	0.773	0.018	-0.001
		0.3	0.524	0.007	0.000	0.2	0.781	0.018	-0.001
30	100	0.0	-0.009	0.000	0.000	0.2	6.834	0.210	-0.007
		0.1	0.517	0.013	0.000	0.2	6.886	0.211	-0.007
		0.2	1.043	0.027	-0.001	0.2	7.003	0.215	-0.007
		0.3	1.558	0.040	-0.001	0.2	7.131	0.219	-0.007
	500	0.0	0.009	0.000	0.000	0.2	6.849	0.210	-0.007
		0.1	0.535	0.014	0.000	0.2	6.928	0.213	-0.007
		0.2	1.059	0.027	-0.001	0.2	7.005	0.215	-0.007
		0.3	1.574	0.040	-0.001	0.2	7.082	0.218	-0.007
	1000	0.0	0.011	0.000	0.000	0.2	6.847	0.210	-0.007
		0.1	0.537	0.014	0.000	0.2	6.924	0.213	-0.007
		0.2	1.061	0.027	-0.001	0.2	6.994	0.215	-0.007
		0.3	1.576	0.040	-0.001	0.2	7.082	0.218	-0.007

Note: G. Mean = Generating Mean; E. Mean = Estimated Mean; G. Var = Generating Variance; E. Var = Estimated Variance

Table 10

The Recovery Analysis for Means and Variances for the 1PL Model

		1PL							
Item	*N*	G. Mean	E. Mean	RMSE	Bias	G. Var	E.Var	RMSE	Bias
10	100	0.0	0.006	0.000	0.000	0.2	0.246	0.001	0.000
		0.1	0.109	0.000	0.000	0.2	0.248	0.002	0.000
		0.2	0.213	0.000	0.000	0.2	0.250	0.002	0.000
		0.3	0.318	0.001	0.000	0.2	0.257	0.002	0.000
	500	0.0	0.009	0.000	0.000	0.2	0.203	0.000	0.000
		0.1	0.110	0.000	0.000	0.2	0.206	0.000	0.000
		0.2	0.210	0.000	0.000	0.2	0.207	0.000	0.000
		0.3	0.311	0.000	0.000	0.2	0.209	0.000	0.000
	1000	0.0	0.009	0.000	0.000	0.2	0.198	0.000	0.000
		0.1	0.109	0.000	0.000	0.2	0.200	0.000	0.000
		0.2	0.209	0.000	0.000	0.2	0.202	0.000	0.000
		0.3	0.310	0.000	0.000	0.2	0.204	0.000	0.000
30	100	0.0	0.009	0.000	0.000	0.2	0.219	0.001	0.000
		0.1	0.110	0.000	0.000	0.2	0.223	0.001	0.000
		0.2	0.212	0.000	0.000	0.2	0.225	0.001	0.000
		0.3	0.314	0.000	0.000	0.2	0.230	0.001	0.000
	500	0.0	0.013	0.000	0.000	0.2	0.198	0.000	0.000
		0.1	0.112	0.000	0.000	0.2	0.201	0.000	0.000
		0.2	0.213	0.000	0.000	0.2	0.203	0.000	0.000
		0.3	0.313	0.000	0.000	0.2	0.205	0.000	0.000
	1000	0.0	0.013	0.000	0.000	0.2	0.196	0.000	0.000
		0.1	0.112	0.000	0.000	0.2	0.199	0.000	0.000
		0.2	0.212	0.000	0.000	0.2	0.201	0.000	0.000
		0.3	0.312	0.000	0.000	0.2	0.202	0.000	0.000

Note: G. Mean = Generating Mean; E. Mean = Estimated Mean; G. Var = Generating Variance; E. Var = Estimated Variance

Table 11

The Recovery Analysis for Mean and Variances for the 2PL Model

		2PL							
Item	*N*	G. Mean	E. Mean	RMSE	Bias	G. Var	E. Var	RMSE	Bias
10	100	0.0	-0.005	0.000	0.000	0.2	0.237	0.001	0.000
		0.1	0.096	0.000	0.000	0.2	0.214	0.000	0.000
		0.2	0.199	0.000	0.000	0.2	0.238	0.001	0.000
		0.3	0.303	0.000	0.000	0.2	0.241	0.001	0.000
	500	0.0	-0.001	0.000	0.000	0.2	0.203	0.000	0.000
		0.1	0.099	0.000	0.000	0.2	0.204	0.000	0.000
		0.2	0.200	0.000	0.000	0.2	0.203	0.000	0.000
		0.3	0.299	0.000	0.000	0.2	0.202	0.000	0.000
	1000	0.0	-0.001	0.000	0.000	0.2	0.200	0.000	0.000
		0.1	0.099	0.000	0.000	0.2	0.200	0.000	0.000
		0.2	0.199	0.000	0.000	0.2	0.200	0.000	0.000
		0.3	0.299	0.000	0.000	0.2	0.200	0.000	0.000
30	100	0.0	-0.005	0.000	0.000	0.2	0.221	0.001	0.000
		0.1	0.096	0.000	0.000	0.2	0.219	0.001	0.000
		0.2	0.197	0.000	0.000	0.2	0.219	0.001	0.000
		0.3	0.300	0.000	0.000	0.2	0.222	0.001	0.000
	500	0.0	-0.001	0.000	0.000	0.2	0.201	0.000	0.000
		0.1	0.099	0.000	0.000	0.2	0.200	0.000	0.000
		0.2	0.199	0.000	0.000	0.2	0.200	0.000	0.000
		0.3	0.300	0.000	0.000	0.2	0.202	0.000	0.000
	1000	0.0	-0.001	0.000	0.000	0.2	0.199	0.000	0.000
		0.1	0.099	0.000	0.000	0.2	0.199	0.000	0.000
		0.2	0.200	0.000	0.000	0.2	0.199	0.000	0.000
		0.3	0.300	0.000	0.000	0.2	0.200	0.000	0.000

Note: G. Mean = Generating Mean; E. Mean = Estimated Mean; G. Var = Generating Variance; E. Var = Estimated Variance

4.2.2 Type I Error Analysis

Table 12 shows the Type I error rates for each of the conditions of the simulation study. The error rates were analyzed at a nominal level of significance of .05. This study assumed control using Bradley's (1978) stringent range of .045 to .055 for a nominal level of .05 for Type I error rates. Thus, error rates less than .045 or greater than .055 were considered loss of Type I error control. The results showed that several conditions were determined to have loss of Type I error control. First, for a test length of 10 and sample size of 100, for of the LGM model, the error rate was .060. Second, when the test length was 10 with N = 500 for the 1PL LIRT-LGM model, the error rate was .060. When the test length = 30 with N = 500 and with N = 1000, the error rates were .065 and .079, respectively, for the LIRT-LGM. Third, when the test length was 30 with N = 1000 of the 2PL LIRT-LGM model, the error rate was .056. High Type I error rates are bolded in Tables 12. As can be seen in Figures 18 and 19, the Type I error rates of the 1PL were worse than those of the other models, in particular, when test length was 30 with sample sizes of N = 500 and 1000. The Type I error rates of the 2PL LIRT-LGM, however, displayed better results compared to the other models. Overall, the Type I error performed well for controlling and setting aside the problematic conditions.

Table 12

The Type I Error for the LGM and LIRT-LGM Models

		LGM	LIRT-LGM	
			1PL	2PL
Item	*N*	Type I Error rate	Type I Error rate	Type I Error rate
10	100	**.060**	.054	.054
	500	.054	**.060**	.049
	1000	.052	.048	.050
30	100	.052	.053	.048
	500	.045	**.065**	.048
	1000	.050	**.079**	**.056**

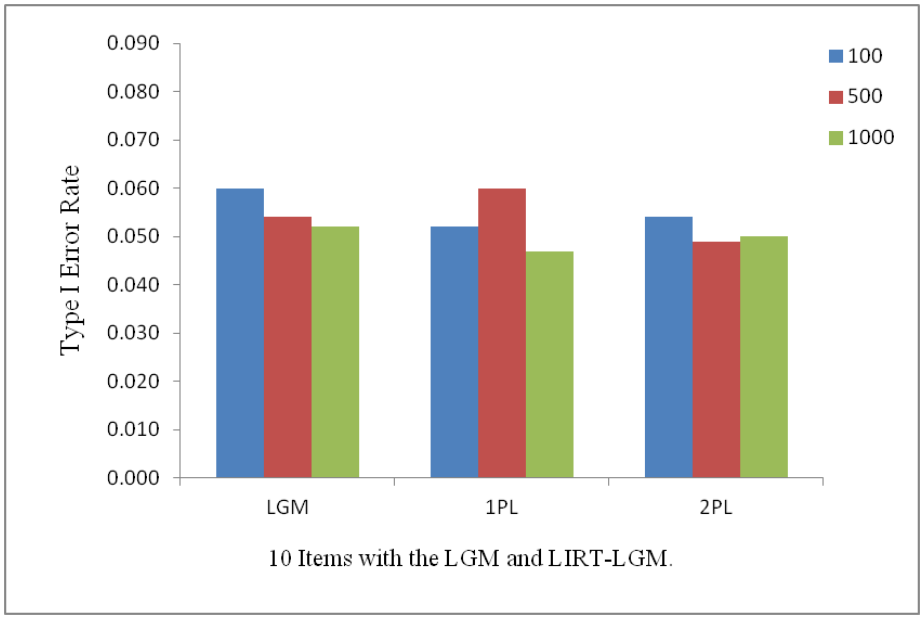

Figure 18. The Type I error rate of the 10 items with the LGM and the LIRT-LGM models.

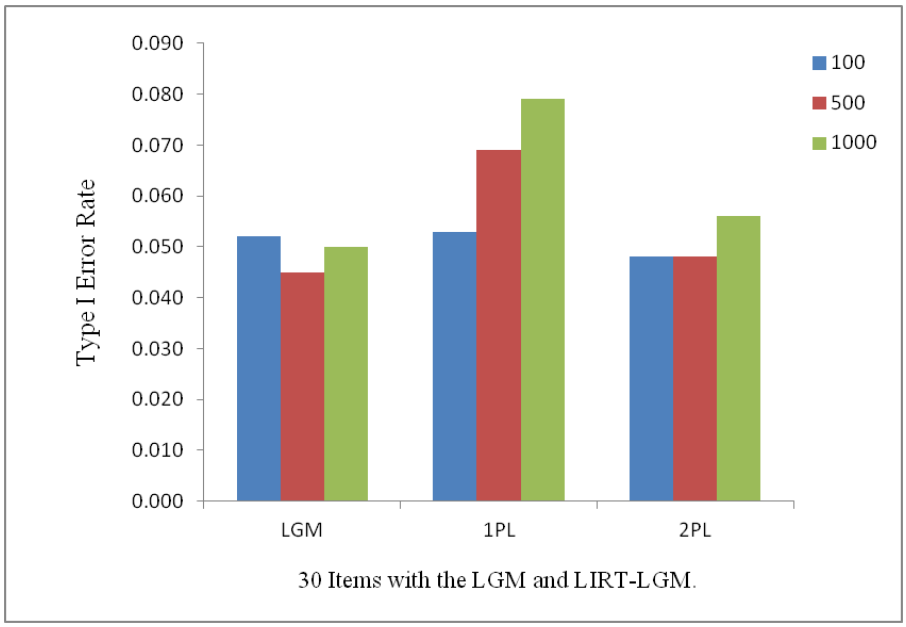

Figure 19. The Type I error rate of the 30 items with the LGM and the LIRT-LGM models.

4.2.3 Power Analysis

The Type I errors were generally controlled for both the LGM and LIRT-LGM models except where noted. Power rates are only reported for those conditions in which the Type I Errors were controlled (see Table 13). Acceptable values for power are usually taken to be at least .80 (Muthén, 2002). Power rates less than .80 are shown in bold in Table 13 and in parentheses in the sequel. For the LGM model, low power rates were observed for several conditions: the 10-item test in the N = 500 sample size with ES = .1 (power = .622), the 30-item test for the N = 100 sample with ES = .1 (power = .191) and ES = .2 (power = .567), and for the 30-item test for the N = 500 sample size with ES = .1 (power =.665).

For the 1PL version of the LIRT-LGM model, low power rates were observed for the 10-item test in the N = 100 sample with ES = .1 (power = .200) and ES = .2 (power =.548). In addition, for the 30-item test in the N = 100 sample with ES = .1 (power =.253) and ES = .2

(power =.626), the power rates also were low. For the 2PL version of the LIRT-LGM model, low power rates were observed for the 10-item test in the N = 100 sample for ES =.1 (power =.177) and ES = .2 (power =.522), and also for the 10-item test in the N = 500 sample for ES = .1 (power =.639). Further, for the 30-item test in the N = 100 sample for ES = .1 (power =.182) and ES = .2 (power =.583), and for the 30-item test in the N = 500 for ES = .1 (power =.690), the power rates were low. In summary, the low power rates were observed for the both LGM and LIRT-LGM model when ES was small (i.e., ES = .1 and .2) for samples of size N = 100 and 500. However, when sample sizes were large (e.g., N = 1000), both the LGM and LIRT-LGM showed high power rates.

4.2.4 Model Selection for the 1PL and 2PL

This study adopts -2loglikelihood to determine the goodness of fit shown in Table 14. The fit indices provided by M*p*lus 7.4 indicated that the 2PL LIRT-LGM provided a good fit to the data since the values of the -2loglikelihood were smaller than the 1PL LIRT-LGM.

Table 13

The Power for the LGM and LIRT-LGM Models

			LGM	LIRT-LGM	
				1PL	2PL
Item	*N*	Effect Sizes	Power rate	Power rate	Power rate
10	100	.1	N/A	**.200**	**.177**
		.2	N/A	**.548**	**.522**
		.3	N/A	.865	.847
	500	.1	**.622**	N/A	**.639**
		.2	.992	N/A	.995
		.3	1.000	N/A	1.000
	1000	.1	.887	.942	.897
		.2	1.000	1.000	1.000
		.3	1.000	1.000	1.000
30	100	.1	**.191**	**.253**	**.182**
		.2	**.567**	**.626**	**.583**
		.3	.868	.920	.905
	500	.1	**.665**	N/A	**.690**
		.2	.997	N/A	.998
		.3	1.000	N/A	1.000
	1000	.1	.922	N/A	N/A
		.2	1.000	N/A	N/A
		.3	1.000	N/A	N/A

Table 14

The Fit Indices for the Model Selection of the 1PL and 2PL

		LIRT-LGM	
		1PL	2PL
Item	*N*	-2loglihood	-2loglihood
10	100	2937.468	2914.848
	500	14753.594	14668.296
	1000	29517.896	29354.842
30	100	8060.984	7979.132
	500	40462.312	40168.140
	1000	80963.516	80403.900

CHAPTER 5

DISCUSSION

Recently, longitudinal research methods have been used to measure individual growth over time; thus, longitudinal data analysis has become more accessible in handling the problem of growth. While analyzing longitudinal data, researchers must make certain that the same constructs are measuring across different time points. Latent growth modeling is commonly used for analyzing change over time. The LGM model describes changes across individuals in the means and variance of the intercept and slope and the covariance of the intercept and slope over time. The unique aspect of the LGM model is combining the individual and group levels of analysis.

The scale construction of the LGM is based on classical test theory so, for this model, the composite score used in the LGM is the sums of the individual items reported as a single score. The advantage of using CTT is that it provides a simple useful means of obtaining a score. One important drawback of the LGM model is that it lacks a mechanism for assessing measurement invariance over time points. As a result, there is no mechanism in the model to test whether or not the same construct is being measured over time points. Consequently, the LGM model may not be a very useful model for drawing correct inferences associated with change. Thus, in the current study, we presented a new model that combined a second-order latent growth model and a longitudinal item response theory model to investigate growth over time. The advantages of the combination with an LIRT includes handling measurement invariance, having item and person statistics on the same scale, and using items to discriminate among respondents based on latent

abilities. There were two purposes for this study. First, it compared the performance of the LGM and the LIRT-LGM using empirical data to analyze depressive symptoms. In the example, the LIRT-LGM was used to analyze depressive symptoms using the 1PL and 2PL models. Second, this study provided a simulation study with different sample sizes, test lengths, and effect sizes to assess Type I errors and power. Results are summarized below for the empirical example and the simulation study. In addition, the suggestions for future research are presented.

5.1 Summary and Discussion

This empirical study used in the example presented in this dissertation included data for three time points of data collected for the Family and Community Health Study (FACHS, Simons et al., 2011). The DSM-IV specifies nine depressive symptoms that were analyzed based on 381 African-American adolescent girls. The results of the empirical example showed that mean depression scores increased over the three-time points. Depressive symptoms increased gradually on all three occasions except for two items, decreased interest or pleasure and suicidality. Results for the LGM model found that the model fit the data. Further, the mean and variance of the intercept, the mean and variance of the slope, and the covariance of the intercept and slope were statistically significant, indicating depression symptoms increased and also that individual differences in increase of depressive symptoms were detected.

Results for the LIRT-LGM, however, were different. The mean of the intercept and of the slope, the covariance of the intercept and slope, and the variance of intercept were not significant, suggesting that the symptoms of depression did not change over time for this sample. Further, the variance of the slope of the 1PL was not significant, meaning no individual differences in trajectories were detected. The variance of the slope of the 2PL model, however,

showed the same results as the LGM, that is, that there was significant variance in the slope. This suggested that individual difference were present in the trajectories of depression. Thus, results for the 1PL LIRT-LGM differed from results for the 2PL LIRT-LGM.

For the simulation study, two test lengths, three sample sizes, and four effect sizes were manipulated. Correlations, mean RMSEs, and mean biases were calculated for evaluating recovery of item parameters. Correlations for both the LGM and LIRT-LGM models indicated clear linear relationships with generating values. The small mean RMSEs and mean bias values for the LIRT-LGM indicated good recovery of generating parameters. Results for the LGM model were less accurate as the mean RMSEs were high. In addition, the recovery analysis of the means and variances of the two LIRT-LGM models were also well recovered. Recovery of the means and variances of the LGM model, however, was poor, since the mean RMSE and mean bias were larger than the LIRT-LGM model. Overall, the recovery analysis of item parameters indicated that estimation algorithms were able to recover the thresholds (or item difficulty) and loading (or item discrimination) for different test lengths and sample sizes. The recovery of parameter estimates from the LIRT-LGM model was better than for the LGM model for each condition.

Type I error rates were analyzed at a nominal level of .05, and the acceptable range in this study was taken to be between .045 and .055. Type I errors were controlled for most conditions. Lack of control was found in one condition for the LGM, three conditions for the 1PL LIRT-LGM, and one condition of the 2PL LIRT-LGM.

This study only reported power when the Type I error was controlled. When the *ES* were .1 and .2 with a sample size of $N = 100$ and $ES = .1$ with sample size of $N = 500$, results indicated

low power for both the LGM and LIRT-LGM models. However, when sample sizes and effect sizes increased, higher power was observed.

Overall, the results of this finding were able to respond to the five research questions: (1). Do depressive symptoms change in girls' early adolescence? Is this change heterogeneous? (2). Are there differences between the performance of the LGM and the LIRT-LGM? (3). Does the 2PL model fit better than the 1PL model? (4). How do the test lengths and sample sizes, influence the Type I error? and (5). How do the test lengths, sample sizes, and effect sizes influence power? In the discussion, this dissertation mainly responded to the research questions and was focused on the statistical performance of both the LGM and LIRT-LGM models.

First, the empirical study found inconsistent results for the LGM and LIRT-LGM. The mean of the LGM model indicated that depressive symptoms increased across three-time points; however, results for the LIRT-LGM suggested that depressive symptoms did not change across three-time points. In addition, the variances of the LGM model and the 2PL version of the LIRT-LGM model indicated that there was heterogeneity around the growth parameter, meaning individual differences were present in the growth trajectories. Based on the empirical results, both methods displayed different results. Thus, a simulation study was used to investigate the Type I error control and power of the two models.

Second, accounting for measurement invariance is important for longitudinal data as researchers need to make certain that the same construct is measured over time. If the invariance does not hold, the results are difficult if even possible to interpret. The major differences between the LGM and LIRT-LGM models are that the LIRT-LGM is able to measure invariance and to have person- and item- statistics on the same scale, but the LGM model does not have these features. Thus, compared to the performance of the LGM and LIRT-LGM models, the

recovery analysis of means and variances of the LGM model were poor recovered. In particular, the estimated values of variances for the 30-item test were different from the generating values. Additionally, the mean RMSE values for this condition were high, indicating less estimation accuracy. Even though the recovery was poor, Type I errors were generally controlled, except for the 10-item test in the N = 100 samples. Recovery was good of thresholds, loadings, means, and variances of the LIRT-LGM models for different test lengths and sample sizes. In this study, the performances of the LIRT-LGM models were better than the LGM model.

Third, -2loglikelihood was used to inform model selection. The result of -2loglikelihood for 2PL LIRT-LGM was smaller than the 1PL LIRT-LGM with all simulation conditions. In addition to fit indices of -2loglikelihood, the recovery analysis of the mean RMSE and mean bias of the 2PL model was recovered better than the 1PL model, since these values were small with all simulation conditions. Thus, results of this study suggest that the 2PL LIRT-LGM fit the simulated data better than the 1PL LIRT-LGM.

Fourth, test lengths and sample sizes do affect the Type I Error rates in both methods. The minimum sample sizes (N=100) of this dissertation were taken from the recommendations for SEM with normal distributions of continuous variables (Anderson & Gerbing, 1988; Newsom, 2015). For the LGM model, when test lengths were increased from 10 to 30, and sample sizes were larger, the Type I Error rates were controlled. For example, when test length was 30 items for the large, N = 1000 sample size, Type I error rates reached a nominal level of significance of .05. The 1PL LIRT-LGM, however, performed poorly when items and sample sizes were large. For instance, for the 30-item test in the N = 500 and 1000 samples, Type I error rates of the 1PL LIRT-LGM were not controlled (i.e., Type I error rates = .069 and .079, respectively, for the two sample sizes). Likewise, for the 2PL LIRT-LGM model, for the 30-item

test for the N = 1000 sample size, the Type I error rate was not controlled (i.e., Type I error rate = .056). In sum, this finding suggested that test lengths and sample sizes were important factors influencing Type I error control.

Fifth, when effect sizes were small (i.e., ES = .1 and .2) with N = 100, the results showed low power for the LIRT-LGM model. However, for a larger ES of .3, the results showed higher power for the LIRT-LGM model. In addition, when effect size was .1 with N = 500, the results showed low power for the LGM and 2PL LIRT-LGM model. Nonetheless, for a ES of .2 witn N = 500, the results showed higher power for both the LGM and LIRT-LGM models. The results also showed high power when N = 1000. Results in Table 13 indicate that, when sample sizes and effect sizes increased, the power rates increased as well. Consequently, test lengths had less effect on, although different sample sizes and effect sizes did for both models.

In addition to different test lengths, sample sizes, and effect sizes, the number of time points would also affect power. Three occasions, at a minimum, are required for detecting a standard linear growth without constraints. Fan and Fan (2005) found that there was a convergence problem with only three-time points. However, there was no convergence problem with small sample size (i.e., fewer than 100 cases) for five or more time-points (Newsom, 2015).

Testing statistical power for the LGM model, it is important to distinguish between tests of the significance of the fixed effects (i.e., mean intercept and slope) and the random effect (i.e., variance intercept and slope). For the testing mean of the intercept or slope, at least a sample size of 100 is needed when effect sizes are small and three time points are involved. Fan and Fan (2005) suggested 50-75 cases with medium effect sizes and 100 cases for a smaller effect sizes for three time points. When testing the variance of the intercept or slope, however, large sample sizes and more time points (e.g., five-time points) are needed. Muthén and Curran (1997) suggest

that for measuring the variance of the slope, sample sizes should be at least 500 when effect sizes are small and five time points are included. In addition, Rast and Hofer (2014) suggest that at least 3,000 cases would be needed for three-time points, at least 1,800 cases for four-time points, and at least 750 cases for five-time points, to have sufficient power when measuring variance of the slope and intercept.

This dissertation followed the previous studies (Fan & Fan, 2005; Maas & Hox, 2004, 2005) and employed three-time points and samples from a minimum of 100 cases to measure sufficient power. The results showed that when *ES* = .1 or .2 and *N* = 100, the power was less than 0.8 for both models for all test lengths studied. In other words, power was not affected by test lengths, although it was affected by sample size and effect size. Consequently, this finding suggests that one needs to employ a minimum of three time points for testing the mean of the slope, but also that *ES* should be at least.3 for a sample of *N* = 100 to have sufficient power.

5.2 Future Research

The results of this study suggested that the LIRT-LGM model performed better than the LGM model under the conditions considered in the simulation study. The recovery analysis was good recovered, Type I errors were controlled, and power was better for the LIRT-LGM models compared to the LGM. Additional manipulation of various conditions would be useful for assessing the performance of the LIRT-LGM model. An important limitation of the empirical data set is that it only investigated depressive symptoms in adolescent African-American girls. Other ethnicities, such as white, Hispanic, and Asian should also study. Second, the growth models can be useful in examining whether there is a significant change in the mean level of test scores over time and whether those shapes show linear or nonlinear growth (Curran & Bollen,

2001). However, this model cannot be used as a diagnosis tool. In a future study, it would be interesting to incorporate the LIRT-LGM model into diagnostic classification modeling (DCM) to diagnose people as depressed. Third, both the empirical and simulation studies only employed three-time points. Including at least four-time points would provide a better opportunity to observe growth. Further, the model would not be limited to just linear change, and could even possibly be extended to a dual change or a triple change model. Fourth, this study only manipulated short test lengths of 10 and 30 items and small effect sizes of .0, .1, .2, and .3. Additional manipulation of these factors, such as longer test lengths (e.g., 50) and medium effect sizes (e.g., .5) or large effect sizes (e.g., .8) might be useful for evaluating the performance of the LIRT-LGM model. Fifth, the LIRT-LGM can be extended to more complex models, such as a LIRT- latent mixture growth model (LIRT-LMGM) in the future. Sixth, polytomous type should also be considered. Seventh, the current study only generated a linear growth pattern and included only complete data. Curvilinear trajectory conditions and missing data should also be considered. Using other estimation algorithms, such as Bayesian estimation, might be useful for exploring the performance of the LIRT-LGM. Eighth, the LIRT-LGM model analyzed data in both empirical and stimulation study used 1PL and 2PL. The three parameter logistic model also can be included in the LIRT-LGM model. Ninth, the average value of the -2 Log likelihood (-2LL) across 1000 replications was used to inform the model selection of the 1PL and 2PL for the LIRT-LGM model in the current study. However, the model fit test using -2LL on each time and a total of repeated 1000 times can also be considered in the future. This would allow one to calculate a percentage of best fit for a specific model.

Overall, according to the results presented, this study suggests that the LIRT-LGM provides more consistent results under a variety of practical testing conditions compared to the

LGM model. The only problematic condition for the LIRT-LGM model was for the short 10-item tests in the small sample size (N = 100) and small effect sizes (ES = 0.1 and 0.2) conditions. Low power was observed for these conditions. Nevertheless, for larger sample sizes, effect sizes, and test lengths, the LIRT-LGM model performs well. The LIRT-LGM appears to be a good choice for analyzing categorical data used in longitudinal research based on the results of the empirical study and simulation study, although it does require larger sample sizes.

REFERENCES

Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1-14). Dordreeht, AH: Springer Netherlands.

Andersen, E. B. (1985). Estimating latent correlations between repeated testing. *Psychometrica*, *50*, 3-16. doi:10.1007/BF02294143

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

Baker, F. B. (2001). *The basic of item response theory*. New York, NY: Eric Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: Taylor & Francis.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York, NY: Academic Press.

Beach, S. R. H., Lei, M-K, Brody, G. H., Simons, R. L., Cutrona, C., & Philibert, R. A. (2012). Genetic moderation of contextual effects on negative arousal and parenting in African-American parents. *Journal of Family Psychology, 26*, 46–55. doi:10.1037/a0026236

Belden, A. C., Pagliaccio, D., Murphy, E. R., Luby, J. L., & Barch, D. M. (2015). Neural activation during cognitive emotion regulation in previously depressed compared to healthy children: Evidence of specific alterations. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*, 771-781.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, *Psychometrika*, 46, 443-459.

Bock, R. D., & Mislevy, R. J. (1981). An item response curve model for matrix-sampling data: The California grade-three assessment. *New Directions for Testing and Measurement*, *10*, 65-90.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models a synthesis of two traditions. *Sociological Methods & Research*, *32*, 336-383. doi: 10.1177/0049124103260222

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). Hoboken, NJ: John Wiley & Sons.

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

Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, *62*, 23-43.

Burchinal, M. R., Nelson, L., & Poe, M. (2006). Growth curve analysis: An introduction to various methods for analyzing longitudinal data. In K. McCartney, M. R. Burchinal, & K. L. Bub (Eds.), *Best practices in quantitative methods for developmentalists* (pp. 65-87). Boston, MA: Blackwell.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*, 421-483.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi: 10.1207/S15328007SEM0902_5

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika, 78*, 576-600. doi: 10.1007/S11336-012-9314-0

Choi, J., Harring, J. R., & Hancock, G. R. (2009). Latent growth modeling for logistic response functions. *Multivariate Behavioral Research, 44*, 620-645. doi: 10.1080/00273170903187657

Cohen, A. S., Bottge, B. A., & Wells, C. S. (2001). Using item response theory to assess effects of mathematics instruction in special populations. *Exceptional Children, 68*, 23-44.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods*

for the analysis of change (pp. 107–135). Washington, DC: American Psychological Association.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*, 121-136. doi: 10.1080/15248371003699969

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*, 295-311.

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, *3*, 196-204.

Duncan, T. E., & Duncan, S. C. (2004). An introduction to latent growth curve modeling. *Behavior Therapy*, *35*, 333-363.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concept, issues, and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*, 74-96. doi: 10.1080/15427600902911163

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495-515. doi:10.1007/BF02294487

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. *Objective Measurement: Theory into Practice*, *4*, 223-236.

Embretson, S. E., & Reise, S. P. (2000*). Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X., & Fan, X. (2005). Power of latent growth modeling for detecting linear growth: Number of measurements and comparison with other analytic approaches. *The Journal of Experimental Education*, *73*, 121-139. doi: 10.3200/ JEXE.73.2.121-139

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

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374. doi:10.1016/0001-6918(73)90003-6

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3–26.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*, 599–624.

Fischer, G. H. (1997). Structural Rasch models: Some theory, applications, and software. *Objective Measurement: Theory into Practice*, *4*, 185-207.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *222*, 309-368.

Fleming, C. B., Mason, W. A., Mazza, J. J., Abbott, R. D., & Catalano, R. F. (2008). Latent growth modeling of the relationship between depressive symptoms and substance use during adolescence. *Psychology of Addictive Behaviors*, *22*, 186-197. doi: 10.1037/0893-164X.22.2.186

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC medicine*, *13*, 1-11. doi: 10.1186/s12916-015-0325-4

Ge, X., Kim, I. J., Brody, G. H., Conger, R. D., Simons, R. L., Gibbons, F. X., & Cutrona, C. E. (2003). It's about timing and change: Pubertal transition effects on symptoms of major depression among African American youths. *Developmental Psychology*, *39*, 430-439. doi: 10.1037/0012-1649.39.3.430

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First-versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 479-503. doi: 10.1080/10705511.2013.797832

Ghisletta, P., & McArdle, J. J. (2012). Latent curve models and latent change score models estimated in R. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 651-682. doi: 10.1080/10705511.2012.713275

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*, 544-565. doi: 10.1207/S15328007SEM1004_4

Hamagami, F., & McArdle, J. J. (2007). Dynamic extensions of latent difference score models. In S. M. Boker & M. J. Wenger (Eds.). *Data analytic techniques for dynamical systems* (pp. 47-85). Mahwah, NJ: Psychology Press.

Hambleton, R. K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage publications.

Hammond, S. (2006). Using psychometric tests. In G. M. Breakwell, S. Hammond, C. Fife-Schaw, & J. A. Smith (Eds.), *Research methods in psychology* (3rd ed.) (pp. 182-209). Thousand Oaks, CA: Sage.

Han, K. T. (2009). IRTEQ: Windows application that implements item response theory scaling and equating. *Applied Psychological Measurement*, *33*, 491-493.

Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 117-159). Greenwich, CT: Information Age.

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth modeling to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 309-341). Greenwich, CT: Information Age.

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

Hardy, S. A., & Thiels, C. (2009). Using latent growth curve modeling in clinical treatment research: An example comparing guided self-change and cognitive behavioral therapy treatments for bulimia nervosa. *International Journal of Clinical and Health Psychology*, *9*, 51-71.

Hayward, C., Killen, J. D., Wilson, D. M., Hammer, L. D, Lift, I. F., & Kraemer, H. C. (1997). Psychiatric risk associated with early puberty in adolescent girls. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 255–262.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.

Hertzog, C., von Oertzen, T., Ghisletta, P., & Linfenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 541-563. doi: 10.1080/10705510802338983

Hofer, S. M., Thorvaldsson, V., & Piccinin, A. M. (2012). Foundational issues of design and measurement in developmental research. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 3-16). New York, NY: Guilford.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modelling. *Family Science Review*, *11*, 354-373.

Hsieh, C.-A., von Eye, A., Maier, K., Hsieh, H.-J., & Chen, S.-H. (2013). A unified latent growth curve model. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 592-615. doi: 10.1080/10705511.2013.824778

Jones, M. C., & Bayley, N. (1950). Physical maturing among boys as related to behavior. *Journal of Educational Psychology*, *41*, 129–148.

Jöreskog, K., & Sörbon, D. (1996). *LISREL 8*: *User's Reference Guide* (2nd ed.). Chicago, IL: Scientific Software International.

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. doi:10.1080/10705510701758406

Keller, P. S., & El-Sheikh, M. (2011). Latent change score modeling of psychophysiological data: An empirical instantiation using electrodermal responding. *Psychophysiology*, *48*, 1577-1586. doi: 10.1111/j.1469-8986.2011.01225.x

Keenan, K., Hipwell, A., Feng, X., Babinski, D., Hinze, A., Rischall, M., & Henneberger, A. (2008). Subthreshold symptoms of depression in preadolescent girls are stable and predictive of depressive disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*, 1433-1442. doi: 10.1097/CHI.0b013e3181886eab

Kline, R. B. (1998). Software review: Software programs for structural equation modeling: Amos, EQS, and LISREL. *Journal of Psychoeducational Assessment*, *16*, 343-364.

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

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling, 14*, 581-610. doi: 10.1080/10705510701575438

LeMoult, J., Ordaz, S. J., Kircanski, K., Singh, M. K., & Gotlib, I. H. (2015). Predicting first onset of depression in young girls: Interaction of diurnal cortisol and negative life events. *Journal of Abnormal Psychology*, *124*, 850-859.

Lindhorst, T., & Oxford, M. (2008). The long-term effects of intimate partner violence on adolescent mothers' depressive symptoms. *Social Science & Medicine*, *66*, 1322-1333.

Lord, F. M. (1952). *A theory of test scores* (Psychometric monographs No. 7). Iowa City, IA: Psychometric Society.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127-137.

Maas, C. J., & Hox, J. J. (2005). Sufficient samples sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86-92.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32,* 215-253. doi: 10.1207/s15327906mbr3203_1

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160

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

Maxwell, S. E., & Tiberio, S. (2007). Multilevel models of change: Fundamental concepts and relationships to mixed models and latent growth-curve models. In A. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 439–452). New York, NY: Oxford University Press.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York, NY: Plenum.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 341–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577-605.

McArdle, J. J., & Bell, R. Q. (2000). An introduction to latent growth models for developmental data analysis. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data* (pp. 69-107). Mahwah, NJ: Lawrence Erlbaum Associates.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110-133.

McArdle, J. J., & Grimm, K. J. (2010). Five steps in latent curve and latent change score modeling with longitudinal data. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 245-273). New York, NY: Springer Heidelberg.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*, 126–149. doi:10.1037/a0015857

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 139-175). Washington, DC: American Psychological Association.

McArdle, J. J., & Hamagami, F. (2004). Methods for dynamic change hypotheses. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 295-335). Dordrecht, AA: Kluwer Academic.

McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen & H.W. Reese (Eds.), *Life-span developmental psychology*: *Methodological innovations* (pp. 223-267). Hillsdale, NJ: Lawrence Erlbaum Associates.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.

McArdle, J. J., Petway, K. T., & Hishinuma, E. S. (2014). IRT for growth and change. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 435-456). New York, NY: Taylor & Francis.

McGirr, A., Renaud, J., Seguin, M., Alda, M., Benkelfat, C., Lesage, A., & Turecki, G. (2007). An examination of DSM-IV depressive symptoms and risk for suicide completion in major depressive disorder: A psychological autopsy study. *Journal of Affective Disorders*, *97*, 203-209.

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

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107-122. doi:10.1007/BF02294746

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*, 5-9.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. O. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, *74*, 807-811.

Muthén, B. O. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*, 43-65.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Pychometrika*, *49*, 115-132.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371-402.

Muthén, L. K. (2002). Using M*p*lus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models (M*p*lus Web Notes, No. 1). Retrieved from https://www.statmodel.com/download/webnotes/mc1.pdf .

Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York, NY: Routledge.

O' Kearney, R., Kang, K., Christensen, H., & Griffiths, K. (2009). A controlled trial of a school-based Internet program for reducing depressive symptoms in adolescent girls. *Depression and Anxiety*, *26*, 65-72.

Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, *18*, 267-296.

Pearson, K. (1927). Foreword. In K. Pearson (Ed.), *Tracts for computers* (pp. iii-viii). London England: Cambridge University Press.

Petersen, A. C., Sarigiani, P. A., & Kennedy, R. E. (1991). Adolescent depression: Why more girls? *Journal of Youth and Adolescence, 20*, 247–271. doi: 10.1007/BF01537611

Pettit, J. W., Roberts, R. E., Lewinsohn, P. M., Seeley, J. R., & Yaroslavsky, I. (2011). Developmental relations between perceived social support and depressive symptoms

through emerging adulthood: Blood is thicker than water. *Journal of Family Psychology, 25*, 127-136. doi: 10.1037/a0022320

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

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics, 14*, 1-17.

Rast, R, & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: Simulation results based on actual longitudinal studies. *Psychological Methods, 19,* 133—154. doi: 10.1037/a0034524

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*, 95-101.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National institute of mental health diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381-389.

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.

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Placentini, J., Davies, M., Conners, C. K., & Regier, D. (1993). The diagnostic interview schedule for children-revised version (DISC-R): I. Preparation, field testing, interrater reliability, and acceptability. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 643-650.

Simons, R. L., Lei, M. K., Beach, S. R. H., Brody, G. H., Philibert, R. A., Gibbons, F. X. (2011). Social environment, genes, and aggression: Evidence supporting the differential susceptibility perspective. *American Sociological Review*, *76*, 883-912. doi: 10.1177/0003122411427580

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72-101.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408. doi:10.1007/BF02294363

Tavares, H. R., & Andrade, D. F. (2006). Item response theory for longitudinal data: Item and population ability parameters estimation. *Test*, *15*, 97-123.

Tisak, J., & Meredith, W. (1990). Descriptive and associative development models. In A. von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. II, pp. 387-406). San Diego, CA: Academic Press.

Titman, A. C., Lancaster, G. A., & Colver, A. F. (2016). Item response theory and structural equation modelling for ordinal data: Describing the relationship between KIDSCREEN and Life-H. *Statistical Methods in Medical Research*, *25*, 1892-1924. doi: 10.1177/0962280213504177

Tucker, L.R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, *23*, 19-23.

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

Wackerly, D. D., Mendenhall, W., & Scheffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont, CA: Thomson Higher Education.

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thiessen & H. Wainer (Eds.), *Test scoring* (pp. 23-72). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23,* 455-465. DOI: 10.1080/10705511.2015.1096744

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Norwell, MA: Kluwer Academic.

Wilson, M., Zheng, X., & McGuire, L. W. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13,* 1–22.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20,* 557-585.

APPENDIX A

The List of Questions of Depressive Symptoms

Depression Manual: Diagnostic Interview Schedule for Children, Version 4 for Time 1 to Time 3

Diagnostic Symptoms	Absent	Present
1. **Depressed mood or irritable** most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful)		
2. **Decreased interest or pleasure** in most activities, most of each day		
3. **Significant weight change (5%) or change in appetite**		
4. **Change in sleep**: Insomnia or hypersomnia		
5. **Change in activity**: Psychomotor agitation or retardation		
6. **Fatigue or loss of energy**		
7. **Guilt/worthlessness:** Feelings of worthlessness or excessive or inappropriate guilt		
8. **Concentration**: diminished ability to think or concentrate, or more indecisiveness		
9. **Suicidality:** Thoughts of death or suicide, or has suicide plan		

APPENDIX B

Threshold (or Item Difficulty Parameter) of Generating and Estimated Parameters for LIRT-LGM with 10 Items

	1PL				2PL			
	Generating Parameter *b*	Estimated Parameter *b*			Generating Parameter *b*	Estimated Parameter *b*		
Item		*N*=100	*N*=500	*N*=1000		*N*=100	*N*=500	*N*=1000
1	-2.0	-2.081	-2.083	-2.083	-2.0	-1.983	-1.994	-1.995
2	-1.5	-1.581	-1.583	-1.583	-1.5	-1.487	-1.495	-1.496
3	-1.0	-1.081	-1.083	-1.083	-1.0	-0.990	-0.997	-0.997
4	-0.5	-0.581	-0.583	-0.583	-0.5	-0.493	-0.498	-0.498
5	0.0	-0.081	-0.083	-0.083	0.0	0.004	0.001	0.001
6	0.0	-0.081	-0.083	-0.083	0.0	0.004	0.001	0.001
7	0.5	0.419	0.417	0.417	0.5	0.500	0.500	0.500
8	1.0	0.919	0.917	0.917	1.0	0.997	0.999	0.998
9	1.5	1.419	1.417	1.417	1.5	1.494	1.497	1.497
10	2.0	1.919	1.917	1.917	2.0	1.991	1.996	1.996

APPENDIX C

Loading (or Item Discrimination Parameter) of Generating and Estimated Parameters for LIRT-LGM with 10 Items

	2PL			
	Generating Parameter *a*	Estimated Parameter *a*		
Item		*N*=100	*N*=500	*N*=1000
1	1.0	1.013	1.005	1.004
2	1.1	1.114	1.105	1.105
3	1.2	1.216	1.206	1.205
4	1.3	1.317	1.306	1.306
5	1.4	1.418	1.407	1.406
6	1.5	1.519	1.507	1.506
7	1.6	1.621	1.608	1.607
8	1.7	1.722	1.708	1.707
9	1.8	1.823	1.809	1.808
10	1.9	1.925	1.909	1.908

APPENDIX D

Threshold (or Item Difficulty Parameter) of Generating and Estimated Parameters for LIRT-LGM with 30 Items

	1PL				2PL			
	Generating Parameter *b*	Estimated Parameter *b*			Generating Parameter *b*	Estimated Parameter *b*		
Item		*N*=100	*N*=500	*N*=1000		*N*=100	*N*=500	*N*=1000
1	-2.0	-2.091	-2.092	-2.092	-2.0	-1.999	-1.996	-1.996
2	-1.5	-1.591	-1.592	-1.592	-1.5	-1.502	-1.497	-1.497
3	-1.0	-1.091	-1.092	-1.092	-1.0	-1.004	-0.998	-0.998
4	-0.5	-0.591	-0.592	-0.592	-0.5	-0.507	-0.499	-0.498
5	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
6	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
7	0.5	0.409	0.408	0.408	0.5	0.489	0.500	0.500
8	1.0	0.909	0.908	0.908	1.0	0.986	0.999	1.000
9	1.5	1.409	1.408	1.408	1.5	1.484	1.498	1.499
10	2.0	1.909	1.908	1.908	2.0	1.981	1.997	1.999
11	-2.0	-2.091	-2.092	-2.092	-2.0	-1.999	-1.996	-1.996
12	-1.5	-1.591	-1.592	-1.592	-1.5	-1.502	-1.497	-1.497
13	-1.0	-1.091	-1.092	-1.092	-1.0	-1.004	-0.998	-0.998
14	-0.5	-0.591	-0.592	-0.592	-0.5	-0.507	-0.499	-0.498
15	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
16	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
17	0.5	0.409	0.408	0.408	0.5	0.489	0.500	0.500
18	1.0	0.909	0.908	0.908	1.0	0.986	0.999	1.000
19	1.5	1.409	1.408	1.408	1.5	1.484	1.498	1.499
20	2.0	1.909	1.908	1.908	2.0	1.981	1.997	1.999
21	-2.0	-2.091	-2.092	-2.092	-2.0	-1.999	-1.996	-1.996
22	-1.5	-1.591	-1.592	-1.592	-1.5	-1.502	-1.497	-1.497
23	-1.0	-1.091	-1.092	-1.092	-1.0	-1.004	-0.998	-0.998
24	-0.5	-0.591	-0.592	-0.592	-0.5	-0.507	-0.499	-0.498
25	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
26	0.0	-0.091	-0.092	-0.092	0.0	-0.009	0.000	0.001
27	0.5	0.409	0.408	0.408	0.5	0.489	0.500	0.500
28	1.0	0.909	0.908	0.908	1.0	0.986	0.999	1.000
29	1.5	1.409	1.408	1.408	1.5	1.484	1.498	1.499
30	2.0	1.909	1.908	1.908	2.0	1.981	1.997	1.999

APPENDIX E

Loading (or Item Discrimination Parameter) of Generating and Estimated Parameters for LIRT-LGM with 30 Items

	2PL			
	Generating Parameter *a*	Estimated parameter *a*		
Item		*N*=100	*N*=500	*N*=1000
1	1.0	1.010	1.003	1.002
2	1.1	1.111	1.104	1.103
3	1.2	1.212	1.204	1.203
4	1.3	1.313	1.304	1.303
5	1.4	1.414	1.405	1.403
6	1.5	1.515	1.505	1.504
7	1.6	1.616	1.605	1.604
8	1.7	1.717	1.706	1.704
9	1.8	1.818	1.806	1.804
10	1.9	1.919	1.906	1.905
11	1.0	1.010	1.003	1.002
12	1.1	1.111	1.104	1.103
13	1.2	1.212	1.204	1.203
14	1.3	1.313	1.304	1.303
15	1.4	1.414	1.405	1.403
16	1.5	1.515	1.505	1.504
17	1.6	1.616	1.605	1.604
18	1.7	1.717	1.706	1.704
19	1.8	1.818	1.806	1.804
20	1.9	1.919	1.906	1.905
21	1.0	1.010	1.003	1.002
22	1.1	1.111	1.104	1.103
23	1.2	1.212	1.204	1.203
24	1.3	1.313	1.304	1.303
25	1.4	1.414	1.405	1.403
26	1.5	1.515	1.505	1.504
27	1.6	1.616	1.605	1.604
28	1.7	1.717	1.706	1.704
29	1.8	1.818	1.806	1.804
30	1.9	1.919	1.906	1.905

APPENDIX F

M*p*lus Code Used for Generating Data with 10 and 30 Items

```
TITLE: Sample sizes with 100 of 10 or 30 dichotomous items were simulated.

 MONTECARLO:
   ! Names used in generating 10 items;
   NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
           u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
           u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

   ! Names used in generating 30 items;
   ! NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
        u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
        u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
        u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
        u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
        u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
        u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
        u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
        u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

! GENERATE used in generating 10 items;
GENERATE = u11 - u103 (1);

   ! GENERATE used in generating 30 items;
   !   GENERATE = u11 - u303 (1);

! CATEGORICAL used in generating 10 items;
CATEGORICAL = u11 - u103;

   ! CATEGORICAL used in generating 30 items;
   ! CATEGORICAL = u11 - u303;

   NOBSERVATIONS = 100;  ! The NOBSERVATIONS can be replaced by 500 and 1000;
   NREPS = 1000;
   SEED = 45335;
   REPSAVE = ALL;
   SAVE = I10N100s_*.DAT;
```

```
 ANALYSIS:  ESTIMATOR = ML;

 Model population:
! USED in generating 10 items;
! (1) to (10) constraint measurement invariance and these show 2PL, as well;

theta1 by u11@1.0 (1)
          u21*1.1 (2)
          u31*1.2 (3)
          u41*1.3 (4)
          u51*1.4 (5)
          u61*1.5 (6)
          u71*1.6 (7)
          u81*1.7 (8)
          u91*1.8 (9)
          u101*1.9 (10);
theta2 by u12@1.0 (1)
          u22*1.1 (2)
          u32*1.2 (3)
          u42*1.3 (4)
          u52*1.4 (5)
          u62*1.5 (6)
          u72*1.6 (7)
          u82*1.7 (8)
          u92*1.8 (9)
          u102*1.9 (10);
 theta3 by u13@1.0 (1)
          u23*1.1 (2)
          u33*1.2 (3)
          u43*1.3 (4)
          u53*1.4 (5)
          u63*1.5 (6)
          u73*1.6 (7)
          u83*1.7 (8)
          u93*1.8 (9)
          u103*1.9 (10);

! USED in generating 30 items;
!theta1 by u11@1.000
               u21*1.111  (1)
               u31*1.212  (2)
               u41*1.313  (3)
               u51*1.414  (4)
               u61*1.515  (5)
               u71*1.616  (6)
```

```
                u81*1.717  (7)
                u91*1.818  (8)
                u101*1.919 (9)
                u111*1.010 (10)
                u121*1.111 (11)
                u131*1.212 (12)
                u141*1.313 (13)
                u151*1.414 (14)
                u161*1.515 (15)
                u171*1.616 (16)
                u181*1.717 (17)
                u191*1.818 (18)
                u201*1.919 (19)
                u211*1.010 (20)
                u221*1.111 (21)
                u231*1.212 (22)
                u241*1.313 (23)
                u251*1.414 (24)
                u261*1.515 (25)
                u271*1.616 (26)
                u281*1.717 (27)
                u291*1.818 (28)
                u301*1.919 (29);

!theta2 by u12@1.000
                u22*1.111(1)
                u32*1.212 (2)
                u42*1.313 (3)
                u52*1.414 (4)
                u62*1.515 (5)
                u72*1.616 (6)
                u82*1.717 (7)
                u92*1.818 (8)
                u102*1.919 (9)
                u112*1.010 (10)
                u122*1.111 (11)
                u132*1.212 (12)
                u142*1.313 (13)
                u152*1.414 (14)
                u162*1.515 (15)
                u172*1.616 (16)
                u182*1.717 (17)
                u192*1.818 (18)
                u202*1.919 (19)
                u212*1.010 (20)
```

```
                u222*1.111 (21)
                u232*1.212 (22)
                u242*1.313 (23)
                 u252*1.414 (24)
                 u262*1.515 (25)
                 u272*1.616 (26)
                 u282*1.717 (27)
                 u292*1.818 (28)
                 u302*1.919 (29);

! theta3 by u13@1.000
                 u23*1.111 (1)
                 u33*1.212 (2)
                 u43*1.313 (3)
                 u53*1.414 (4)
                 u63*1.515 (5)
                 u73*1.616 (6)
                 u83*1.717 (7)
                 u93*1.818 (8)
                 u103*1.919 (9)
                 u113*1.010 (10)
                 u123*1.111 (11)
                 u133*1.212 (12)
                 u143*1.313 (13)
                 u153*1.414 (14)
                 u163*1.515 (15)
                 u173*1.616 (16)
                 u183*1.717 (17)
                 u193*1.818 (18)
                 u203*1.919 (19)
                 u213*1.010 (20)
                 u223*1.111 (21)
                 u233*1.212 (22)
                 u243*1.313 (23)
                 u253*1.414 (24)
                 u263*1.515 (25)
                 u273*1.616 (26)
                 u283*1.717 (27)
                 u293*1.818 (28)
                 u303*1.919 (29);

 [theta1-theta3@0];  ! the mean of the latent factor;
  theta1-theta3@1;    ! the variance of the latent factor;

!set up thresholds
```

```
        [u11$1*-2.0  u12$1*-2.0  u13$1*-2.0]  (11);
        [u21$1*-1.5  u22$1*-1.5  u23$1*-1.5]  (12);
        [u31$1*-1.0  u32$1*-1.0  u33$1*-1.0]  (13);
        [u41$1*-0.5  u42$1*-0.5  u43$1*-0.5]  (14);
        [u51$1*0.0   u52$1*0.0   u53$1*0.0 ]  (15);
        [u61$1*0.0   u62$1*0.0   u63$1*0.0 ]  (16);
        [u71$1*0.5   u72$1*0.5   u73$1*0.5 ]  (17);
        [u81$1*1.0   u82$1*1.0   u83$1*1.0 ]  (18);
        [u91$1*1.5   u92$1*1.5   u93$1*1.5 ]  (19);
        [u101$1*2.0  u102$1*2.0  u103$1*2.0]  (20);

! Item thresholds all estimated with 30 Items;
[u11$1*-1.999     u12$1*-1.999    u13$1*-1.999] (30);
 [u21$1*-1.502     u22$1*-1.502    u23$1*-1.502] (31);
 [u31$1*-1.004     u32$1*-1.004    u33$1*-1.004] (32);
 [u41$1*-0.507     u42$1*-0.507    u43$1*-0.507] (33);
 [u51$1*-0.009     u52$1*-0.009    u53$1*-0.009] (34);
 [u61$1*-0.009     u62$1*-0.009    u63$1*-0.009] (35);
 [u71$1*0.489      u72$1*0.489     u73$1*0.489]  (36);
 [u81$1*0.986      u82$1*0.986     u83$1*0.986]  (37);
 [u91$1*1.484      u92$1*1.484     u93$1*1.484]  (38);
 [u101$1*1.981   u102$1*1.981     u103$1*1.981]  (39);
 [u111$1*-1.999  u112$1*-1.999   u113$1*-1.999] (40);
 [u121$1*-1.502  u122$1*-1.502   u123$1*-1.502] (41);
 [u131$1*-1.004  u132$1*-1.004   u133$1*-1.004] (42);
 [u141$1*-0.507  u142$1*-0.507   u143$1*-0.507] (43);
 [u151$1*-0.009  u152$1*-0.009   u153$1*-0.009] (44);
 [u161$1*-0.009  u162$1*-0.009   u163$1*-0.009] (45);
 [u171$1*0.489   u172$1*0.489    u173$1*0.489]  (46);
 [u181$1*0.986   u182$1*0.986   u183$1*0.986]  (47);
 [u191$1*1.484   u192$1*1.484   u193$1*1.484]  (48);
 [u201$1*1.981   u202$1*1.981   u203$1*1.981]  (49);
 [u211$1*-1.999 u212$1*-1.999  u213$1*-1.999] (50);
 [u221$1*-1.502 u222$1*-1.502  u223$1*-1.502] (51);
 [u231$1*-1.004 u232$1*-1.004  u233$1*-1.004] (52);
 [u241$1*-0.507 u242$1*-0.507  u243$1*-0.507] (53);
 [u251$1*-0.009 u252$1*-0.009  u253$1*-0.009] (54);
 [u261$1*-0.009 u262$1*-0.009  u263$1*-0.009] (55);
 [u271$1*0.489 u272$1*0.489    u273$1*0.489]  (56);
 [u281$1*0.986 u282$1*0.986    u283$1*0.986]  (57);
 [u291$1*1.484 u292$1*1.484    u293$1*1.484]  (58);
 [u301$1*1.981 u302$1*1.981    u303$1*1.981]  (59);
```

```
i s | theta1@0 theta2@1 theta3@2;
i with s*0;
[i@0 s*0];     ! The mean of intercept and slope;
               ! The mean s* can be replaced by 0.1, 0.2, and 0.3 while detecting power;
  i*1  s*0.2;  ! The variance of intercept and slope;
```

APPENDIX G

M*p*lus code for Analyzing LGM with 10 and 30 Items

```
Title: LGM model with 10 or 30 items and 100 N;

 Data: File is F:\Simulation\I10N1002PLsm_list.dat;
       Type=montecarlo;

  Define:

    !the observed variables used in 10 items;
    DEP1=u11+u21+u31+u41+u51+u61+u71+u81+u91+u101;
     DEP2=u12+u22+u32+u42+u52+u62+u72+u82+u92+u102;
     DEP3=u13+u23+u33+u43+u53+u63+u73+u83+u93+u103;

     !the observed variables used in 30 items;
    ! DEP1=u11+u21+u31+u41+u51+u61+u71+u81+u91+u101+u111+
                u121+u131+u141+u151+u161+u171+u181+u191+u201+
                u211+u221+u231+u241+u251+u261+u271+u281+u291+u301;
     !DEP2=u12+u22+u32+u42+u52+u62+u72+u82+u92+u102+u112+
                u122+u132+u142+u152+u162+u172+u182+u192+u202+
                u212+u222+u232+u242+u252+u262+u272+u282+u292+u302;
     !DEP3=u13+u23+u33+u43+u53+u63+u73+u83+u93+u103+u113+
                 u123+u133+u143+u153+u163+u173+u183+u193+u203+
                 u213+u223+u233+u243+u253+u263+u273+u283+u293+u303;

  Variable:
!The variables used in 10 items;
     NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
                   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
                   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

!The variables used in 30 items;
! NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
                  u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
                  u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
```

```
            u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
            u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
            u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
            u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
            u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
            u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;
    USEVARIABLES are DEP1 DEP2 DEP3;

  Analysis: ESTIMATOR IS ML;
            LINK IS LOGIT;

  MODEL:
  !@0, @1, @2 are measurement occasions;
   i s | DEP1@0 DEP2@1 DEP3@2;

! using in 10 Items
   DEP1*2.266;
   DEP2*2.244;
   DEP3*2.445;

! using in 30 Items
     !DEP1*2.499;
     !DEP2*2.443;
     !DEP3*2.715;

   [s*0];     ! The mean of slope;
              ! The mean s* can be replaced by 0.1, 0.2, and 0.3;
    s*0.2;   ! Variance of the slope;

  Output: Tech9;
```

APPENDIX H

M*p*lus Code Used for Analyzing the 1PL Model with 10 Items

```
Title: LIRT-LGM model of Rasch model with 10 items and 100 obs;

 Data: File is F:\Simulation\I10N1002PLsm_list.DAT;
      Type=montecarlo;


 Variable:
   NAMES are
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

   USEVARIABLES are
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

   CATEGORICAL ARE
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

 Analysis:
        ESTIMATOR IS ML;
        PROCESSORS=4;

 Model:
           theta1 by u11-u101* (1);
           theta2 by u12-u102* (1);
           theta3 by u13-u103* (1);
```

```
!Factor mean=0 and variance=1 for identification
    [theta1-theta3@0];
    theta1-theta3@1;

!Item thresholds all estimated
 [u11$1*-2.081   u12$1*-2.081  u13$1*-2.081] (2);
 [u21$1*-1.581   u22$1*-1.581  u23$1*-1.581] (3);
 [u31$1*-1.081   u32$1*-1.081  u33$1*-1.081] (4);
 [u41$1*-0.581   u42$1*-0.581  u43$1*-0.581] (5);
 [u51$1*-0.081   u52$1*-0.081  u53$1*-0.081] (6);
 [u61$1*-0.081   u62$1*-0.081  u63$1*-0.081] (7);
 [u71$1*0.419    u72$1*0.419   u73$1*0.419 ] (8);
 [u81$1*0.919    u82$1*0.919   u83$1*0.919 ] (9);
 [u91$1*1.419    u92$1*1.419   u93$1*1.419 ] (10);
[u101$1*1.919   u102$1*1.919  u103$1*1.919 ] (11);

 i s | theta1@0 theta2@1 theta3@2;

 [i@0 s*0.0];  !Mean;
  i*1  s*0.2;  !Variance;


 output:
   TECH9;
```

APPENDIX I

M*p*lus Code Used for Analyzing the 1PL Model with 30 Items

```
Title: LGM-LIRT model of Rasch model with 30 items and 100 obs;

 Data: File is C:\Simulation\Data\I30N1002PL\I30N1002PLsm_list.DAT;
     Type=montecarlo;


 Variable:
   NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
         u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
         u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
         u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
         u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
         u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
         u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
         u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
         u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

   USEVARIABLES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
        u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
        u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
        u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
        u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
        u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
        u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
        u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
        u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

   CATEGORICAL = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
        u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
        u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
        u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
        u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
```

```
        u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
        u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
        u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
        u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

 Analysis:
        ESTIMATOR IS ML;
        PROCESSORS=4;

 Model:
        theta1 by u11-u301*(1);
   theta2 by u12-u302*(1);
        theta3 by u13-u303*(1);


 !Factor mean=0 and variance=1 for identification
    [theta1-theta3@0];
     theta1-theta3@1;


!Item thresholds all estimated
[u11$1*-2.091  u12$1*-2.091   u13$1*-2.091] (2);
[u21$1*-1.591  u22$1*-1.591   u23$1*-1.591] (3);
[u31$1*-1.091  u32$1*-1.091   u33$1*-1.091] (4);
[u41$1*-0.591  u42$1*-0.591   u43$1*-0.591] (5);
[u51$1*-0.091  u52$1*-0.091   u53$1*-0.091] (6);
[u61$1*-0.091  u62$1*-0.091   u63$1*-0.091] (7);
[u71$1*0.409    u72$1*0.489    u73$1*0.409 ] (8);
[u81$1*0.909    u82$1*0.986    u83$1*0.909 ] (9);
[u91$1*1.409    u92$1*1.484    u93$1*1.409 ] (10);
[u101$1*1.909  u102$1*1.981  u103$1*1.909 ] (11);
[u111$1*-2.091  u112$1*-2.091 u113$1*-2.091] (12);
[u121$1*-1.591  u122$1*-1.591 u123$1*-1.591] (13);
[u131$1*-1.091  u132$1*-1.091 u133$1*-1.091] (14);
[u141$1*-0.591  u142$1*-0.591 u143$1*-0.591] (15);
[u151$1*-0.091  u152$1*-0.091 u153$1*-0.091] (16);
[u161$1*-0.091  u162$1*-0.091 u163$1*-0.091] (17);
[u171$1*0.409   u172$1*0.489  u173$1*0.409 ] (18);
[u181$1*0.909   u182$1*0.986  u183$1*0.909 ] (19);
```

```
[u191$1*1.409   u192$1*1.484  u193$1*1.409 ] (20);
[u201$1*1.909   u202$1*1.981  u203$1*1.909 ] (21);
[u211$1*-2.091  u212$1*-2.091 u213$1*-2.091] (22);
[u221$1*-1.591  u222$1*-1.591  u223$1*-1.591] (23);
[u231$1*-1.091  u232$1*-1.091  u233$1*-1.091] (24);
[u241$1*-0.591  u242$1*-0.591  u243$1*-0.591] (25);
[u251$1*-0.091  u252$1*-0.091  u253$1*-0.091] (26);
[u261$1*-0.091  u262$1*-0.091  u263$1*-0.091] (27);
[u271$1*0.409    u272$1*0.489   u273$1*0.409 ] (28);
[u281$1*0.909   u282$1*0.986   u283$1*0.909 ] (29);
[u291$1*1.409   u292$1*1.484   u293$1*1.409 ] (30);
[u301$1*1.909   u302$1*1.981   u303$1*1.909 ] (31);

i s | theta1@0 theta2@1 theta3@2;

[i@0 s*0.0];  !Mean;
i*1  s*0.2;  !Variance;

output:
TECH9;
```

APPENDIX J

M*p*lus Code Used for Analyzing the 2PL Model with 10 Items

```
Title: LGM-LIRT model of 2PL model with 10 items and 100 obs;

 Data: File is C:\Simulation\I10N1002PLsm_list.DAT;
     Type=montecarlo;


 Variable:
   NAMES are
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

   USEVARIABLES are
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

   CATEGORICAL ARE
   u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
   u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
   u13 u23 u33 u43 u53 u63 u73 u83 u93 u103;

 Analysis:
        ESTIMATOR IS ML;
        PROCESSORS=4;

 Model:
 theta1 by u11@1.000
             u21*1.114 (1)
```

```
            u31*1.216 (2)
            u41*1.317 (3)
            u51*1.418 (3)
            u61*1.519 (5)
            u71*1.621 (6)
            u81*1.722 (7)
            u91*1.823 (8)
            u101*1.925(9);

 theta2 by u12@1.000
            u22*1.114 (1)
            u32*1.216 (2)
            u42*1.317 (3)
            u52*1.418 (3)
            u62*1.519 (5)
            u72*1.621 (6)
            u82*1.722 (7)
            u92*1.823 (8)
            u102*1.925(9);

 theta3 by u13@1.000
            u23*1.114 (1)
            u33*1.216 (2)
            u43*1.317 (3)
            u53*1.418 (3)
            u63*1.519 (5)
            u73*1.621 (6)
            u83*1.722 (7)
            u93*1.823 (8)
            u103*1.925(9);

!Factor mean=0 and variance=1 for identification
    [theta1-theta3@0];
     theta1-theta3@1;

!Item thresholds all estimated
 [u11$1*-1.983  u12$1*-1.983  u13$1*-1.983] (10);
 [u21$1*-1.487  u22$1*-1.487  u23$1*-1.487] (11);
 [u31$1*-0.990  u32$1*-0.990  u33$1*-0.990] (12);
```

```
[u41$1*-0.493  u42$1*-0.493  u43$1*-0.493] (13);
[u51$1*0.004   u52$1*0.004   u53$1*0.004 ] (14);
[u61$1*0.004   u62$1*0.004   u63$1*0.004 ] (15);
[u71$1*0.500   u72$1*0.500   u73$1*0.500 ] (16);
[u81$1*0.997   u82$1*0.997   u83$1*0.997 ] (17);
[u91$1*1.494   u92$1*1.494   u93$1*1.494 ] (18);
[u101$1*1.925  u102$1*1.925  u103$1*1.925] (19);

i s | theta1@0 theta2@1 theta3@2;

 [i@0 s*0.0];  !Mean;
  i*1 s*0.2;  !Variance;

output:
    TECH9;
```

APPENDIX K

M*p*lus Code Used for Analyzing the 2PL Model with 30 Items

```
Title: LGM-LIRT model of 2PL model with 30 items and 100 obs;

 Data: File is C:\Simulation\Data\Type I error\I30N1002PL\I30N1002PLsm_list.DAT;
     Type=montecarlo;


 Variable:
   NAMES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
         u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
         u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
         u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
         u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
         u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
         u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
         u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
         u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

   USEVARIABLES = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
        u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
        u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
        u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
        u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
        u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
        u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
        u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
        u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

   CATEGORICAL = u11 u21 u31 u41 u51 u61 u71 u81 u91 u101
        u111 u121 u131 u141 u151 u161 u171 u181 u191 u201
        u211 u221 u231 u241 u251 u261 u271 u281 u291 u301
```

```
        u12 u22 u32 u42 u52 u62 u72 u82 u92 u102
        u112 u122 u132 u142 u152 u162 u172 u182 u192 u202
        u212 u222 u232 u242 u252 u262 u272 u282 u292 u302
        u13 u23 u33 u43 u53 u63 u73 u83 u93 u103
        u113 u123 u133 u143 u153 u163 u173 u183 u193 u203
        u213 u223 u233 u243 u253 u263 u273 u283 u293 u303;

Analysis:
        ESTIMATOR IS ML;
        PROCESSORS=6;

Model:
   theta1 by u11@1.000
                u21*1.111  (1)
                u31*1.212  (2)
                u41*1.313  (3)
                u51*1.414  (4)
                u61*1.515  (5)
                u71*1.616  (6)
                u81*1.717  (7)
                u91*1.818  (8)
                u101*1.919 (9)
                u111*1.010 (10)
                u121*1.111 (11)
                u131*1.212 (12)
                u141*1.313 (13)
                u151*1.414 (14)
                u161*1.515 (15)
                u171*1.616 (16)
                u181*1.717 (17)
                u191*1.818 (18)
                u201*1.919 (19)
                u211*1.010 (20)
                u221*1.111 (21)
                u231*1.212 (22)
                u241*1.313 (23)
                u251*1.414 (24)
                u261*1.515 (25)
                u271*1.616 (26)
                u281*1.717 (27)
```

```
            u291*1.818 (28)
            u301*1.919 (29);

theta2 by u12@1.000
            u22*1.111 (1)
            u32*1.212 (2)
            u42*1.313 (3)
            u52*1.414 (4)
            u62*1.515 (5)
            u72*1.616 (6)
            u82*1.717 (7)
            u92*1.818 (8)
            u102*1.919 (9)
            u112*1.010 (10)
            u122*1.111 (11)
            u132*1.212 (12)
            u142*1.313 (13)
            u152*1.414 (14)
            u162*1.515 (15)
            u172*1.616 (16)
            u182*1.717 (17)
            u192*1.818 (18)
            u202*1.919 (19)
            u212*1.010 (20)
            u222*1.111 (21)
            u232*1.212 (22)
            u242*1.313 (23)
            u252*1.414 (24)
            u262*1.515 (25)
            u272*1.616 (26)
            u282*1.717 (27)
            u292*1.818 (28)
            u302*1.919 (29);

theta3 by u13@1.000
            u23*1.111 (1)
            u33*1.212 (2)
            u43*1.313 (3)
            u53*1.414 (4)
            u63*1.515 (5)
```

```
            u73*1.616 (6)
            u83*1.717 (7)
            u93*1.818 (8)
            u103*1.919 (9)
            u113*1.010 (10)
            u123*1.111 (11)
            u133*1.212 (12)
            u143*1.313 (13)
            u153*1.414 (14)
            u163*1.515 (15)
            u173*1.616 (16)
            u183*1.717 (17)
            u193*1.818 (18)
            u203*1.919 (19)
            u213*1.010 (20)
            u223*1.111 (21)
            u233*1.212 (22)
            u243*1.313 (23)
            u253*1.414 (24)
            u263*1.515 (25)
            u273*1.616 (26)
            u283*1.717 (27)
            u293*1.818 (28)
            u303*1.919 (29);

!Factor mean=0 and variance=1 for identification
   [theta1-theta3@0];
    theta1-theta3@1;

!Item thresholds all estimated
[u11$1*-1.999 u12$1*-1.999   u13$1*-1.999] (30);
[u21$1*-1.502 u22$1*-1.502   u23$1*-1.502] (31);
[u31$1*-1.004 u32$1*-1.004   u33$1*-1.004] (32);
[u41$1*-0.507 u42$1*-0.507   u43$1*-0.507] (33);
[u51$1*-0.009 u52$1*-0.009   u53$1*-0.009] (34);
[u61$1*-0.009 u62$1*-0.009   u63$1*-0.009] (35);
[u71$1*0.489  u72$1*0.489    u73$1*0.489]  (36);
[u81$1*0.986  u82$1*0.986    u83$1*0.986]  (37);
[u91$1*1.484  u92$1*1.484    u93$1*1.484]  (38);
[u101$1*1.981 u102$1*1.981  u103$1*1.981]  (39);
```

```
[u111$1*-1.999  u112$1*-1.999  u113$1*-1.999] (40);
[u121$1*-1.502  u122$1*-1.502  u123$1*-1.502] (41);
[u131$1*-1.004  u132$1*-1.004  u133$1*-1.004] (42);
[u141$1*-0.507  u142$1*-0.507  u143$1*-0.507] (43);
[u151$1*-0.009  u152$1*-0.009  u153$1*-0.009] (44);
[u161$1*-0.009  u162$1*-0.009  u163$1*-0.009] (45);
[u171$1*0.489   u172$1*0.489   u173$1*0.489]  (46);
[u181$1*0.986   u182$1*0.986   u183$1*0.986]  (47);
[u191$1*1.484   u192$1*1.484   u193$1*1.484]  (48);
[u201$1*1.981   u202$1*1.981   u203$1*1.981]  (49);
[u211$1*-1.999  u212$1*-1.999  u213$1*-1.999] (50);
[u221$1*-1.502  u222$1*-1.502  u223$1*-1.502] (51);
[u231$1*-1.004  u232$1*-1.004  u233$1*-1.004] (52);
[u241$1*-0.507  u242$1*-0.507  u243$1*-0.507] (53);
[u251$1*-0.009  u252$1*-0.009  u253$1*-0.009] (54);
[u261$1*-0.009  u262$1*-0.009  u263$1*-0.009] (55);
[u271$1*0.489   u272$1*0.489    u273$1*0.489]  (56);
[u281$1*0.986   u282$1*0.986    u283$1*0.986]  (57);
[u291$1*1.484   u292$1*1.484    u293$1*1.484]  (58);
[u301$1*1.981   u302$1*1.981   u303$1*1.981]  (59);

i s | theta1@0 theta2@1 theta3@2;

[i@0 s*0.0];  !Mean;
 i*1  s*0.2;  !Variance;

output:
   TECH9;
```