An Explanatory Item Response Model Approach for Studying Latent Growth in Alphabet Knowledge

---

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Xiaoxin (Elizabeth) Wei, M.S., B.S.

August, 2015

Abstract

This study reviews growth modeling techniques, and then focused specifically on the use of explanatory item response models for studying growth while accounting for lack of time-invariance among item properties. Using this framework, results suggested there was a significant amount of growth in kindergarteners' alphabet knowledge from fall to spring of a school year. Individual differences in latent ability and growth were large initially, but became considerably smaller by the end of the year. The difficulty of different item properties influenced examinee's responses and the person properties have significant impact on the amount of latent growth. Implications were discussed from both substantive and methodological perspectives.

Key words: growth, explanatory item response model, lack of invariance, alphabet recognition, letter sounds

Dedication

I dedicate this dissertation to my family, my advisor, and a few close friends who have been helping me and supporting me all along my precious and adventurous PHD journey.

Acknowledgements

I would like to appreciate the greatest support from my families, without which I would not be able to complete my doctoral degree.

My advisor and committee chair, J. Patrick Meyer, has provided paramount support, guidance, and patience which are indispensible to me to complete my doctoral study and this dissertation.

My committee members, Tim Konold, Marcia Invernizzi, and Ji Hoon Ryoo have also provided essential support and help along the entire process of my dissertation.

I am grateful!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

The purpose of the present study is to review the literature on growth models with a particular focus on latent growth explanatory item response models (LG-EIRMs). It applies a LG-EIRM that accounts for time-varying item parameters, investigates latent growth in alphabet knowledge, and examines the impact of certain child properties on growth. Item responses to the lower-case alphabet recognition subtest and the letter sounds subtest of Phonological Awareness Literacy Screening-Kindergarten (PALS-K) were studied and analyzed for growth. These two subtests are two required tasks of PALS-K that measure a child's performance in the alphabet knowledge construct (Invernizzi et al., 2011). To be more specific, the lower-case alphabet recognition subtest measures a child's ability to provide names of all lower-case alphabet letters. It includes 26 binary items representing 26 lower-case alphabet letters. The letter sounds subtest measures a child's ability to produce sounds associated with individual letters. It includes 26 binary items as well.

Data comprise item-level responses of examinees to the lower-case alphabet recognition subtest and the letter sounds subtest that were measured on three occasions, fall, mid-year, and spring of the 2013-2014 school year. The sample used in the study consisted of 5,000 examinees randomly selected from a large sample of kindergarteners from a mid-Atlantic state who took PALS-K test in 2013-2014 school year.

Growth Modeling

Castellano and Ho (2013a) defined growth broadly as "the academic performance of a student or group over two or more time points." Briggs & Betebenner (2009) stated growth is reflected by changes in student achievement over time and typical questions about growth concern the magnitude of growth and adequacy of growth. Growth studies utilize statistical methods to examine student achievement data and model growth over different time points. Different statistical models are available for studying growth.

Old approaches of studying growth, such as ANOVA and multiple regression techniques, only analyze the average change of the variable of interest between different time points and ignore individual differences. To capture information about individual differences in growth researchers may employ a different set of statistical methods such as multilevel growth models, linear mixed models, and student growth percentiles (SGP; Betebenner & Linn, 2009). Most of these methods focus on observed scores and fail to recognize the role of measurement error, but methods that focus on latent growth overcome this limitation. One such approach involves structural equation modeling (SEM). The involvement of SEM creates an effective way to take measurement error into consideration and study growth on the latent construct, instead of the observed score. Using SEM techniques, growth curve models (GCMs) provide the capability to account for measurement errors and to examine not only within-person growth over time but also between-person variability in the within-person growth. One limitation of GCMs is that they are designed for continuous outcome variables (Finney & DiStefano, 2006) which are usually observed test scores, instead of item scores. They can be adapted for categorical data through use of polychoric correlations (Muthén, 1983; Muthén, 1984).

However, Item Response Theory (IRT) provides a more direct approach for studying latent growth with categorical data.

IRT refers to a conceptual framework for studying the relationship between the probability of a categorical response outcome and examinee ability and item characteristics (Lord & Novick 1968; van der Linden & Hambleton, 1997; Embretson & Reise, 2000). Traditional unidimensional IRT models are descriptive (DeBoeck & Wilson, 2004) in that they include a parameter for person ability (i.e. the latent trait or measured construct) and one or more parameters for item characteristics such as difficulty or discrimination. Building upon descriptive models, generalized linear mixed models (GLMM) and nonlinear mixed models (McCulloch & Searle, 2001) allow person properties, item properties, or both item and person properties to be included in an item response model. As such, Wilson and DeBoeck (2004) refer to these methods as explanatory item response models (EIRMs). A major feature of EIRM is the combination of measurement of person ability and estimation of research design factors (DeBoeck & Wilson, 2004, p.26). EIRM is able to accommodate research designs based on between-person factors, such as person groups defined by certain person properties (e.g. gender, intervention group, or race). It can also be used to incorporate within-person design factors, such as measurement occasions and item properties (Wilson & Moore, 2011).

If only item properties are included, the EIRM becomes a linear logistic test model (LLTM) which uses item properties to explain differences between item difficulties (Fischer, 1973; DeBoeck & Wilson, 2004, p.61). Instead of modeling individual item's contribution to person response, the LLTM models the effect of each

item property on an item response. It also allows for modeling interactions between item properties. If only person properties are included in the model, the EIRM becomes a latent regression model which regress the latent trait on person properties, such as gender, race, or disability status. As a person explanatory model, latent regression models the effect of person properties on an item response (Verhelst & Eggen, 1989; DeBoeck & Wilson, 2004, p58). When both item and person properties are included, the EIRM becomes the latent regression LLTM, a doubly explanatory model which allows for estimation of effect of both item and person properties on response. Each of these EIRMs focuses on measurement at a single time point, but they can be extended to account for longitudinal data and examinee growth.

Within the conceptual framework of IRT, a variety of models have been developed to study longitudinal growth (Andersen, 1985; Embretson, 1991; Wilson, Zheng, & McGuire, 2013; Pastor & Beretvas, 2006). Some of those models have been extended to longitudinal EIRMs that incorporate person and item properties (Cho et al., 2013; Wilson, Zheng, & McGuire, 2013; Stevenson et al., 2013), which are referred as latent growth explanatory item response model (LG-EIRM) in this study. The LG-EIRMs provide a flexible set of tools for studying growth and combining the measurement and explanatory phases into a single model. LG-EIRMs can also be specified to incorporate interactions between person and item properties or between time and item properties (DeBoeck & Wilson, 2004; Wilson, Zheng, & McGuire, 2013). This study used LG-EIRM to study latent growth in alphabet knowledge.

Early Literacy and Alphabet Knowledge

In recent years, increasing interest has focused on literacy development among kindergarteners and younger children (Missall & McConnell, 2010). It is hypothesized that preschool years are a crucial period for educators to apply strategies to shape and positively impact a child's literacy growth trajectories (VanDerHeyden, Snyder, Broussard, & Ramsdell, 2008). Early literacy assessment data allows a child's learning progress to be monitored, and as a result, early literacy growth studies have become important. There are two primary benefits from early literacy growth studies. One is to obtain information about the relation between early reading performance and later reading achievement (Dickinson, Tabors, & Roach, 1996). Another benefit is to explore and understand the influences of different factors on literacy growth of children. The information about sources of individual differences in growth can guide customization of instructional support provided by teacher or school.

Alphabet knowledge is one of the core components of early literacy. Invernizzi et al. (2004) wrote that early literacy mainly consists of four components: phonological awareness, alphabet knowledge, concepts of word, and grapheme-phoneme correspondence. Among those, alphabet knowledge refers to a child's knowledge and ability to recognize all respects of written letters, including letter forms, letter names, and letter sounds (Huang, Tortorelli, & Invernizzi, 2014). Multiple studies have indicated that alphabet knowledge is one of the most powerful predictors of later reading performance (Adams, 1990; Foulin, 2005; Hammill, 2004; Stevenson & Newman, 1986). Additionally, Scarborough (1998) conducted a comparison study and found that alphabet knowledge

could be as powerful a predictor of future reading performance as an entire literacy assessment.

As the two essential components of alphabet knowledge, alphabet recognition refers to a child's ability to identify the names of letters given their corresponding graphic shapes (Evans et al., 2006) and letter sounds is defined as a child's ability to "provide the sounds associated with a particular letter form," (Huang et al., 2014). A child's performance in alphabet recognition and letter sounds is affected by letter-specific features, such as visual confusability (i.e. shape confusability) and letter-name structure. Shape confusability is defined as the shape similarity of the letter to other letters (or numbers) (Huang & Invernizzi, 2014) and it has impact on child's ability of naming different letters. Letter-name structure is determined by the phonological relationship between letter sounds and their letter names (Huang & Invernizzi, 2012; McBride-Chang, 1999) and it influences how child learn to identify the sounds of different letters.

Research Questions

To better understand how a child develops alphabet knowledge, this study examined latent growth in alphabet knowledge (consisting of letter names and letter sounds) within one school year. Factors associated with individual differences in growth and various item properties were also investigated. Person properties included pre-k schooling, English language learner status, disability status, and a child's age in fall. Item properties included letter-shape visual confusability and letter-name structure. Given that item parameter estimates do not necessarily stay constant across groups or test occasions (Embretson & Reise, 2000; Rupp & Zumbo, 2006) this study also evaluated of the

tenability of assuming time-invariant item properties and adjusted the model whenever this assumption was not feasible. Therefore, this study addressed the following research questions:

1. What is the amount of average latent growth in alphabet recognition and letter sounds over three time points?
2. What examinee properties (e.g. pre-k schooling, English Language Learner status) influence the latent growth?
3. Does any item property (i.e. letter-shape confusability and letter-name structure) show a lack of time-invariance?

Method

This study constructed multiple LG-EIRMs based on an IRT-based latent growth model developed by Embretson (1991) and its extensions proposed by Wilson, Zheng, and McGuire (2013). The foundational model of this study was Embretson's (1991) Multidimensional Rasch Model for Learning and Change (MRMLC). Her MRMLC includes an initial latent trait dimension representing initial status and one or more latent growth dimension representing the change between successive time points. The dimensions are assumed to be correlated and item difficulty parameters of the same item are assumed to be invariant across time points in the model. In my study, item properties (i.e. letter-shape confusability and letter-name structure) and person properties (e.g. pre-K schooling and English language learner status) were included in the LG-EIRMs to explain the influence from those item- and person-associated factors. All constructed LG-EIRMs (see Chapter 3) were compared and model fit to the data was evaluated.

Implications

From a methodological perspective, this study provided insight into the benefits and challenges of using explanatory IRT approach to analyze categorical response data and examine the growth of latent constructs. The approach utilized by this study was able to identify person properties that significantly influence the latent growth. Additionally, it built upon Embretson's MRMLC and demonstrated the way to test for time-invariance of item properties or how to incorporate time-varying item properties if necessary.

With respect to early literacy, the results of this study promoted understandings of kindergartener's developmental growth in alphabet knowledge. The findings will shed light on alphabet knowledge curricular planning and instructional design. The relationship between item properties and kindergartener's performance can be taken advantage by teachers to understand kindergartener's learning progress of letters with different particular features and make more effective use of instructional time. Moreover, children are heterogeneous group (Cabell, Justice, Konold, & McGinty, 2011). Understanding how growth in alphabet knowledge is attributable to different person properties benefits future differentiated curricula design and customized intervention for children with specific characteristics.

Chapter 2 Literature Review

A variety of growth modeling techniques, including item response theory-based growth modeling approach, are available to be utilized to examine measurement data and estimate academic growth. The availability of longitudinal data of early literacy assessment allows for growth models to investigate child's developmental growth and monitor a child's learning progress of early literacy core skills. This chapter gives a review on growth modeling and its application in the field of early literacy.

Growth Models

High-stakes testing in education has traditionally focused on measuring student performance at a single point in time (see No Child Left Behind Act of 2001). This type of testing only measures a student's current state of knowledge or achievement *status* but it does not reflect student *growth* – a student's learning or achievement over time (Castellano & Ho, 2013a; Briggs & Betebenner, 2009). Measuring status is easier to accomplish as it requires data collection at a single point in time. Growth measurement, on the other hand, is more difficult to accomplish as it requires data collection at multiple time points with the same or very similar measure on each occasion (Castellano & Ho, 2013a). Complicating matters are the multitude of statistical methods for studying growth. Available methods include gain-score models, multilevel growth models, growth curve models, latent growth models, and student growth percentiles (SGP; Betebenner &

Linn, 2009). They all answer questions related to the magnitude of growth or the adequacy of growth (Briggs & Betebenner, 2009), but each method has its strengths and limitations. A researcher should understand the benefits of each approach and choose a method most suited for the research questions at hand.

Briggs and Betebenner (2009) broadly classified growth models as either relative or absolute growth models. A *relative growth model* evaluates test performance relative to prior achievement, whereas an *absolute growth model* evaluates test performance conditional on time. The main idea of relative growth is borrowed from pediatrics where physicians measure and describe the height and weight of infant in pediatrics in terms of growth percentiles. In education, *student growth percentiles* (SGPs) are the percentile rank of a student's current test score conditional on prior achievement (Betebenner, 2009). SGPs provide a relative (i.e. norm-referenced) interpretation of growth. Linn (2008) believes SGPs are an appropriate and compelling descriptive measure of growth which can serve as the basis of a more systematic and proper educational accountability system. There are multiple approaches to estimating SGPs, but the simplest method is one that calculates percentile ranks from the test scores at the current time point based on examinee test scores from a previous time point. For example, examinees that score in the bottom ten percent on the initial exam are combined into a group, and percentile ranks of the current exam scores are then computed for this group. The process is repeated for students scoring in the second ten percent on the initial exam and so on until there are ten sets of growth percentiles characterizing relative growth for the entire group of examinees. The STAR Assessments (Renaissance Learning, 2012) and AIMS Web (Pearson, 2012b) are two examples of operational educational measures that use this

method. A limitation of this simple approach is that it can only be used to compare two time points (current test score and prior test score; Grady, Lewis, & Gao, 2010). To look at test scores obtained from multiple points in time, you must consider all possible pairings of them.

Betebenner (2009) introduced the use of quantile regression as a more sophisticated method for computing SGPs and studying relative growth. This method allows one to compute percentiles ranks of current exam performance conditional on one or more prior test scores or even conditional on other data such as student gender. Betebenner's method is widely used in high-stakes educational settings such as the Colorado Growth Model Program (Briggs & Betebenner, 2009). Relative growth models provide an interpretive framework that helps test users understand how a student's change in achievement compares to other students who started at a similar level of achievement. They do not provide a way to quantify the actual amount of growth. Absolute growth models overcome this limitation and quantify not only the amount of growth, but also the type of growth (e.g. linear, quadratic).

Examples of absolute growth models include gain-score models, multilevel models, growth curve models, and latent growth models. The gain-score model is the most intuitive approach to quantify the amount of growth, since the gain score is simply the difference in a student's performance on the same test administered at two different time points (Castellano & Ho, 2013a). The gain-score model can be extended to describe the average growth at the group level by taking average of gain scores of a group of examinees. Traditional approaches of studying absolute growth are primarily ANOVA

and multiple regression. These two approaches compute the average change in the variable of interest between different time points and consider it as growth. The main limitation of such early approaches is that they ignore differences among individuals and treat those differences as error variance. Consequently, useful and valuable information about growth hidden in the error variance would not be analyzed.

Random-effects ANOVA, random coefficient modeling, and multilevel modeling (Preacher, Wichman, MacCallum, & Briggs, 2008) allow for individual differences in growth. Specifically speaking, these methods include random coefficients in their models and estimations of those random effects reflect individual differences in growth. In a multilevel model, level-one involves the measures obtained from each time point, and individual variation in initial status and growth at level-two. Multilevel models allow researchers to explore the relationship between person-specific covariates and individual differences in growth (Duncan & Duncan, 2004; Pastor & Beretvas, 2006; Preacher et al., 2008). They have a number of strengths including their ability to capture individual difference in growth via using random coefficients and their well-established statistical estimation procedures (Duncan & Duncan, 2004). They also make it easy to incorporate many time points for each individual and for each individual to have different numbers of time points measured at unequal intervals. A limitation of using a multilevel model to study growth is that it is essentially a univariate method that does not extend easily to multivariate outcomes (Kaplan, 2009).

Growth curve models (GCMs) provide a way to study growth with structural equation modeling techniques. It therefore allows for multiple outcome variables and

ways to examine the relationship between growth and various covariates such as examinee property variables that are consistent over time and environmental variables that change over time (Duncan & Duncan, 2004; Kaplan, 2009). GCMs also provide the capability to examine not only within-person growth over time but also between-person variability in the within-person growth. GCMs and multilevel models are not mutually exclusive methods for studying growth. Indeed, the same model can be constructed from either framework. Raudenbush and Bryk (2002) described a multilevel modeling approach for GCM. A limitation of GCMs is that they are designed for continuous outcome variables (Finney & DiStefano, 2006). They can be adapted for categorical data through use of polychoric correlations (Muthén, 1983; Muthén, 1984). However, item response theory provides a more direct approach for studying growth with categorical data.

Explanatory Item Response Models

Item response theory (IRT) models refer to a set of models for studying the relationship between the probability of a categorical response outcome and examinee ability and item properties (Linden & Hambleton, 1997; Embretson & Reise, 2000). Traditional unidimensional IRT models are descriptive (DeBoeck & Wilson, 2004) in that they include a parameter for person ability and one or more parameters for item characteristics such as difficulty and discrimination. The most basic IRT model is the Rasch model (Rasch, 1960/1980). Given a person's ability θ_p and an item's difficulty, β_i, the probability of person p responding correctly to the item i is given by,

$$P(X_{pi}=1|\theta_p)=\frac{\exp(\theta_p-\beta_i)}{1+\exp(\theta_p-\beta_i)}, \quad (2.1)$$

and the logit of Equation 2.1 is $\text{logit}[P(X_{pi}=1|\theta_p)]=\eta_{pi}=\theta_p-\beta_i$. The difficulty parameter describes the location of an item characteristic curve such that smaller values move the curve to the left and larger values move it to the right. Additional parameters may be included in the model to control the slope of the curve and its lower asymptote. These additions result in the two-parameter logistic and three-parameter logistic models, respectively. Descriptive IRT models also exist for polytomous items. Two examples are the partial credit model (Masters, 1982) and the generalized partial credit model (Muraki, 1992).

Researchers have demonstrated the way IRT can be situated within a generalized linear mixed model (Adams, Wilson, & Wang, 1997; Adams, Wilson, & Wu, 1997; Wu, Adams, & Wilson 1998; Kamata, 2001; Wilson & DeBoeck, 2004) and generalized nonlinear mixed model (Rijmen, Tuerlinckx, DeBoeck, & Kuppens, 2003). Generalized linear mixed models (GLMM) and nonlinear mixed models (NLMM; McCulloch & Searle, 2001) allow person covariates, item covariates, or both item and person covariates to be included in an item response model. As such, Wilson and DeBoeck (2004) refer to these methods as explanatory item response models (EIRMs).

The EIRM is defined as the item response model that seeks to explain item response in terms of its relation to other covariates (DeBoeck & Wilson, 2004). As implied by its name, EIRM uses explanatory approach to reveal the nature of the relation among item response and item-relevant or person-relevant property variables in the

model. Specifically speaking, in the context of item response modeling, EIRM explains and depicts the mathematical relationship among item response, person latent ability, and item characteristics. Depending on the properties included in the model, an EIRM can be explanatory on either person side, item side or both.

A GLMM is a generalized case of linear mixed model that extends to categorical outcomes. A Linear mixed model is basically one type of linear regression model, but with two important unique features. First, the model contains two types of independent variables, variables with fixed weights estimating fixed effects which do not vary as a function of observed individual units and variables with random weights estimating individual-specific random effects which vary across observed individual units. The second feature is that the distribution of individual regression weights follows a specified mathematical format (DeBoeck & Wilson, 2004, p.21). However, generally, linear mixed models can only be applied to continuous outcome data and the error term in the model is continuous which would not recognize the boundaries of categorical variable. The generalized linear mixed model is able to handle categorical data. To be more specific, it involves three parts to connect the categorical observed dependent variable to a combination of independent variables in a linear function: (a) the observed dependent variable Y_{pi} is related to its expected value π_{pi} through an independent Bernoulli distribution, (b) a link function connects the expected value of the categorical dependent variable π_{pi} to the expected value η_{pi} of the underlying continuous variable, and (c) the expected value of the underlying continuous variable η_{pi} is linked to the combination of a number of independent variables in the linear mixed function (see McCullagh &

Nelder, 1989; DeBoeck & Wilson, 2004, p.28). As a result, the categorical observed dependent variable is mathematically related to the independent variables within a linear mixed function.

Equation 2.2 show the most general form of the GLMM for binary data (DeBoeck & Wilson, 2004).

$$\eta_{pi} = \sum_{j=1}^{J} \theta_{pj} Z_{ij} + \sum_{k=0}^{K} \beta_k X_{ik} \,. \tag{2.2}$$

where θ_p are the random effects assumed to have a multivariate normal distribution, $\theta_p \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. The β_k are fixed regression weights, typically representing item characteristics such as item difficulty. It is common practice to multiply the β_k by -1, so that the parameters are interpreted as difficulty and not easiness. With such multiplication, equation 2.2 becomes,

$$\eta_{pi} = \sum_{j=0}^{J} \theta_{pj} Z_{ij} - \sum_{k=0}^{K} \beta_k X_{ik} \,. \tag{2.3}$$

Finally, the vectors Z_i and X_i contain person and item property variables. They can be dummy coded to produce different types of item response models.

Coding Z_p with a vector of ones, the random effect component becomes θ_{p0}, which is viewed as the person ability parameter in the IRT framework. Coding X_i such that $X_{ik} = 1$ when $i = k$ and $X_{ik} = 0$ otherwise results in the Rasch model,

$$\eta_{pi} = \theta_p - \beta_i \tag{2.4}$$

where η_{pi} is the logit of person p's probability of correctly answering item i, θ_p is the latent ability parameter of person p, and β_i is the item difficulty of item i.

Person or item property variables can be coded in Z_i and X_i to model the relationship between different person or item characteristics and the latent ability. For example, a Rasch model with covariates indicating both person properties and item properties is referred to as a doubly explanatory Rasch model in equation 2.5 (DeBoeck & Wilson, 2004) such that the model includes a latent regression for person ability and fixed item effects,

$$\eta_{pi} = \sum_{j=1}^{J} \theta_j Z_{pj} + \varepsilon_p - \sum_{k=0}^{K} \beta_k X_{ik} \qquad (2.5)$$

where ε_p is an error term for person p after all person group property effects are accounted for (i.e. the random person effect). In this doubly explanatory item response model, person and item-relevant property covariates include fixed effects. As for contribution from the person's side, an error term is added to the equation since person ability is usually seen as random effect and the fixed effect of person property may not be able to account for all variances in the item response. On the item side, there is no error term for individual item property or item group attribute effect since contribution from item's side is assumed to be fully explained in terms of individual item parameter or item property effect. However, explanatory Rasch models have variations. Depending on the research interest on certain item or person properties, the basic components of the general formulation 2.5 can be modified.

If only item property covariates are included, the EIRM becomes a linear logistic test model (LLTM) which uses item properties to explain differences between items (Fischer, 1973; DeBoeck & Wilson, 2004, p.61). Unlike Rasch model which models individual item's contribution to person response, the LLTM estimates the effect of item properties on person response and the value of each property for individual items. Particularly, the LLTM also allows for estimation of interaction effect between item properties. DeBoeck and Wilson (2004) present the formulation of LLTM as equation 2.6:

$$\eta_{pi} = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik} \tag{2.6}$$

where X_{ik} is now coded such that it equal 1 if item i involves item property k $(k = 0,...,K)$, and it is zero otherwise. The coefficient β_k is interpreted as the regression weight of item property k. A limitation of the LLTM is that it attempts to account for all of the variation in item difficulty. That is, it makes the strong assumption that those item properties are able to fully explain the item effects (DeBoeck & Wilson, 2004). Janssen and DeBoeck (2006) developed the random weight LLTM (RW-LLTM) by adding a random term to the item side,

$$\eta_{pi} = \theta_p - (\sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i) \tag{2.7}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^{\,2})$. This model is essentially a linear regression on item difficulty such that the covariates in X_i predict difficulties of item properties.

If only person property covariates are included, the EIRM becomes a person explanatory model, a latent regression model which regresses latent trait on a number of external person properties (e.g. gender, race, intelligence). The latent regression Rasch model is given by,

$$\eta_{pi} = \sum_{j=1}^{J} \theta_j Z_{pj} + \varepsilon_p - \beta_i \qquad (2.8)$$

where θ_j is the fixed regression weight of person property j, and ε_p is an error term for person *p* representing the remain effect after accounting for all person properties. As indicated by its formulation, the latent regression model uses person properties to explain differences among person ability.

One major feature of EIRM is the combination of measurement of person ability and estimation of research design factors (DeBoeck & Wilson, 2004, p.26). Since property covariates represent information pertaining to person properties, item properties, or the interactions between them, item response in an EIRM can be linked to information which explains variances in item response. Wilson and Moore (2011) pointed out that a regular IRT model simply provides a description of items and persons via addressing their locations on a common scale, whereas the EIRM explains their locations by using properties of items or persons. Additionally, due to its explanatory nature, EIRM is able to accommodate research designs based on between-person factors, such as person groups defined by certain person properties (e.g. gender, intervention group, or race). It can also be used to incorporate within-person design factors, such as item properties,

since all persons take the same set of items and item properties vary among items within the item set (Wilson & Moore, 2011).

Explanatory Item Response Models for Latent Growth (LG-EIRM)

Andersen (1985) developed a longitudinal growth model for dichotomous response data based on the Rasch model. In his model, the same items are taken by examinees at multiple time points, and the item parameters are assumed to be invariant across time points. His model uses a separate latent trait to describe achievement at each time point, and these time-varying latent traits are assumed to be correlated. Andersen's model does not have a parameter to estimate overall latent growth, but it allows for any type of growth between time points (Wilson et al., 2011). Since the underlying latent trait being measured in the model is essentially the same, the correlations between each time-point-specific latent trait are usually quite strong.

Embretson (1991) developed another Rasch-based growth model referred to as the Multidimensional Rasch Model for Learning and Change (MRMLC). Her model includes at least two dimensions, an initial latent trait dimension representing initial status, and a latent growth dimension representing the change between successive time points. The dimensions are assumed to be correlated and item difficulty parameters of the same item are assumed to be invariant across time points in the model. For an item given at $t = 1,...,T$ time points, the MRMLC is given by,

$$\eta_{pit} = \sum_{t=1}^{T} \theta_t Z_{pt} + \varepsilon_p - \beta_i , \qquad (2.9)$$

where Z_p is a matrix coded with a Weiner simplex process (Embretson,1991). An example of this matrix for three time points is given by,

$$Z_p = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The random effect ε_p is obtained from the first column of Z_p, and it indicates an examinee's initial status on the latent trait. The parameter $\theta_{\cdot 1}$ reflects the change in latent trait from time one to time two, and $\theta_{\cdot 2}$ indicates the amount of change from time two to time three. The Figure 2.1 presents a simple case of MRMLC with two items and three test occasions (see Wilson, Zheng, & McGuire, 2011).

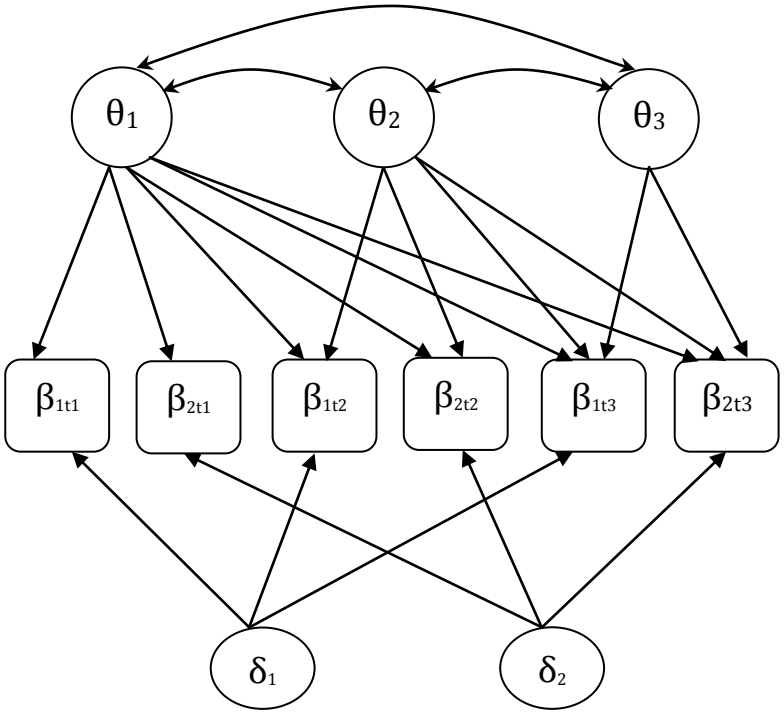

Figure 2.1 An example of Embretson's (1991) MRMLC

Embretson's model can also be extended to link latent growth to changes in cognitive processes and knowledge structures by adding a structural model for item difficulty parameter (Embretson, 1995). It can also be adapted to include additional person property covariates. Stevenson, Hickendorff, Resing, Heiser, and de Boeck (2013) expanded on Embretson's model by incorporating item and person property covariates into the model. Specifically, they added to Embretson's work by incorporating a RW-LLTM with a continuous item property covariate and a random error term for items, instead of fixed individual item effects.

The logit of their model is,

$$\eta_{pit} = \left(\sum_{j=1}^{J} \theta_{jt} Z_{pjt} + \varepsilon_{pt} \right) - \left(\sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \right), \qquad (2.10)$$

where ε_p and ε_i are normally distributed with zero means and variances $\sigma^2_{\varepsilon_p}$ and $\sigma^2_{\varepsilon_i}$. With both time points and person properties included, an example of the matrix Z_{pjt} for three time points with person property j can be shown as,

$$Z_{pjt} = \begin{pmatrix} 1 & 0 & 0 & j_p \\ 1 & 1 & 0 & j_p \\ 1 & 1 & 1 & j_p \end{pmatrix},$$

where j_p represents the value of person property j for person p. In using this model, Stevenson and her colleagues (Stevenson et al., 2013) successively added person and item property covariates to construct a series of nested longitudinal EIRMs. Each model was compared to the previous one in terms of model fit and the model with better fit was

selected. In the best fitting growth model, estimates of the effects of those person properties were used to explain individual differences in growth. In another extension of Embretson's work, Cho, Athay, and Preacher (2013) proposed a Generalized Explanatory Longitudinal Item Response Model (GELIRM) that can be used with multidimensional tests.

Andersen's work, Embretson's MRMLC, and the various extensions provide a flexible set of tools for studying growth. Their main limitation is that the dimensionality increases as the number of time points increases. This feature is not a limitation when examinees are tested on just a few occasions, but it presents notable difficulties or estimation when examinees are observed many times. Adding restrictions to the model can overcome this limitation. For example, the change parameters can be constrained to represent a linear trend across all time points (see Wilson, Zheng, & McGuire, 2013; Pastor & Beretvas, 2006).

All LG-EIRM discussed above assume that item parameters (or item group parameters) are invariant across all time points. If the assumption is incorrect, then estimates of examinee ability and growth will be compromised (Wells, Subkoviak, & Serlin, 2002; Pastor & Beretvas; 2006). The next section discussed parameter invariance in more detail and the subsequent section presents an LG-EIRM that accounts for a lack of invariance.

Item Response Theory and Parameter Invariance

Parameter invariance is a fundamental property of item response theory (IRT). It applies to both item and person parameters. Item parameter invariance refers to the

equality of item parameters, up to a linear transformation, over different samples of examinees. Invariance makes IRT models a popular choice over other measurement models (Rupp & Zumbo, 2006) as it allows for innovative testing procedures such as computerized adaptive testing.

Invariance is a property of the parameters and it is not guaranteed to hold true for parameter estimates. Lack of invariance (LOI) refers to the condition when invariance fails to hold. Invariance must be empirically tested and not just assumed to be true. Item parameter drift (IPD) and differential item function (DIF) are two concepts that reflect the existence of a LOI. Goldstein (1983) defined IPD as the change in an item's parameters over multiple points in time. It may also be considered as the difference in an item's parameters when the item is given to examinees taking different test forms. Differential item function (DIF) is related to this latter concept in that it refers to a difference in an item's parameters when the item is given to different groups of examinees such as male and female test takers (Camili, 2006; Holland & Wainer, 1993; Zumbo, 1999). IPD and DIF are different manifestations of a LOI.

LOI in item parameters affects examinee's response probability and ability scores (Rupp & Zumbo, 2006). Hence, any violation of the parameter invariance property jeopardizes model parameter estimation and person ability score interpretation. The impact of LOI on ability estimates has been investigated and addressed by several studies (Bock, Muraki, & Pfeiffenberger, 1988; Wells, Subkoviak & Serlin, 2002; Wollack, Sung, and Kang, 2005; Wollack, Sung, and Kang, 2006; Han & Guo, 2011), however, results so far are not conclusive in terms of how serious the consequences are. Moreover, since the

invariance property of IRT implies that item parameters are independent of examinee groups, it technically provides the foundation of test fairness. Any type of violation to this property would become a threat to test fairness because it means that an item may be easier for some examinees solely because of their group membership (e.g. race, gender). Hence, LOI should be detected and neutralized before interpreting the test scores and making the inferences from the test results. Or, the response model should not assume invariance and include parameters that account for it.

Item parameter invariance is more complex in growth models as it must hold true in multiple ways. Broadly speaking in longitudinal studies, invariance is defined as the "stability in the psychometric properties of a measure across populations or occasions." (Mellenbergh, 1994; Meredith & Millsap, 1992). More specifically, Cho, Athay, and Preacher (2013) note that invariance in growth models refers to the stability of item parameters: (a) across person groups, (b) across time points (drift), (c) across person groups within a time point, and (d) across time points within a person group. Any violation of invariance at item level impacts person ability estimation such as a person's initial status and growth (Wells, Subkoviak, & Serlin, 2002; Pastor & Beretvas; 2006; Rupp & Zumbo, 2006). To eliminate or minimize the impact of any violation of measurement invariance, two options are typically available. The first one is to detect any lack of invariance before analyzing the growth and drop the items with serious LOI effect (see Cho, Athay, & Preacher, 2013; Kim & Camilli, 2014). The second one is to modify the growth model by incorporating time-by-item or group-by-item interactions (see Paster & Beretvas, 2006; Wilson, Zheng, & McGuire, 2011). Significant interaction effects would indicate a LOI. This study will use the latter approach.

Accounting for LOI in an LG-EIRM

LOI can be incorporated into a LG-EIRM through various interaction terms. A person-by-item interaction represents the difference in difficulty for various person groups (i.e. DIF). It is derived as the product of a person group membership indicator and an item indicator (DeBoeck & Wilson, 2004). A LG-IRM with a DIF effect is given by,

$$\eta_{pit} = \left(\sum_{j=1}^{J} \theta_{jt} Z_{pjt} + \varepsilon_p \right) - \left(\sum_{k=1}^{K} \left(\beta_k X_{ik} + \delta_k W_{pik} \right) + \varepsilon_i \right) \qquad (2.11)$$

where δ_k is the person group-by-item interaction effect, and W_{pik} is the indicator of interaction effect which equals to the product of person group indicator and an item indicator. If the DIF effects, δ_k, are not statistically significant they can be eliminated from the model and the result is Embretson's MRMLC (Equation 2.9).

Extending the application of interaction effect to longitudinal EIRM, a time-by-item interaction term can be used to model lack of invariance of item parameters across different time points (i.e. drift). The LG-EIRM with this type of interaction can be used to test the time invariance of item parameters or item properties while analyzing growth. It is given by

$$\eta_{pit} = \left(\sum_{j=1}^{J} \theta_{jt} Z_{pjt} + \varepsilon_p \right) - \left(\sum_{k=1}^{K} \left(\beta_k X_{ik} + \gamma_k V_{tik} \right) + \varepsilon_i \right) \qquad (2.12)$$

where γ_k is the time-by-item interaction effect, and V_{tik} is the indicator of interaction effect which equals to the product of the time-point indicator and an item indicator X_i. Equation 2.12 is very similar to Equation 2.11. The only difference is the nature of the

interaction term. Like Equation 2.11, if the drift effects γ_k, are not statistically significant or the model with the interaction terms does not improve the overall fit to the data, they can be eliminated from the model and the result is also Embretson's MRMLC (Equation 2.9).

I used a series of LG-EIRMs to study growth in early literacy skills and test for item parameter invariance. This approach is well-suited for early literacy measures because of what the research says about item properties in measures of early literacy and what is known about the development of early literacy skills. The next section provides a brief review of early literacy research. It focuses on information that is aligned with my research questions and is therefore included in various LG-EIRMs in this study.

Estimation Methods of LG-EIRM

Maximum likelihood (ML) estimation is the most widely applied method for estimating item parameters of item response models. All approaches of ML estimation are applied under the assumption that person ability levels are unknown and both item and person parameters have to be estimated from the same response data. Depending on how unknown person abilities are handled in estimation process, three popular approaches of ML are available.

Among them, Joint Maximum Likelihood (JML) estimation uses provisional person ability levels as known person ability parameter values at the beginning of estimation. Then the estimated person parameters will be used to improve the provisional person ability parameters by replacing them. It is an iterative procedure which involves

multiple estimations of persona and item parameters. In the GLMM framework, JML approach treats person ability parameters as fixed effects, the same way as it views item parameters (DeBoeck & Wilson, 2004, p.344). JML has its strengths, such as that it is easily programmable and computationally efficient. In addition, it can be applied to various IRT models although it is typically applied to Rasch model and its extensions. However, its major limitation is the inconsistency of item parameter estimates for fixed length texts because the item parameter estimates from using JML change as sample size changes. In addition, it also produces biased item parameter estimates and does not provide parameter estimates of items with perfect scores (Embretson & Reise, 2000).

Compared to JML, the Conditional Maximum Likelihood (CML) approach is more restrained since it requires that a sufficient statistic for estimating person abilities is available in the data (Embretson & Reise, 2000, p. 215). So CML can only be applied to Rasch model and its extensions, such as partial credit model and rating scale model. In CML, the sum score, instead of person ability parameter, is used to express response pattern probability. It involves an iterative procedure to search for the item parameter estimate that maximizes the response pattern likelihood. In the GLMM framework, CML approach does not need person-specific effects to estimate item parameters or item-specific effects (DeBoeck & Wilson, 2004, p.345).An important strength of CML is it factors out the person ability parameters during estimation and as a result, the item parameter estimates are not influenced by the person ability distribution of the sample. However, the limitations of CML include its inability to provide estimates for item or person with perfect scores, its restrained applicability and loss of information due to

maximization of the conditional likelihood (Embreston & Reise, 2000; DeBoeck & Wilson, 2004, p.345).

The Marginal Maximum Likelihood (MML) approach treats probabilities of response patterns as expectations from a population distribution and views the observed response data as a random sample of a population (Bock & Lieberman, 1970). Bock and Aitkin (1981) developed the Expectation-Maximization (EM) algorithm for MML to estimate item parameters which is also an iterative procedure. In MML estimation for GLMM, person-specific effects (i.e. person ability parameters) are treated as random effects. The vector of unknown population parameters that describes the characteristic of the person ability distribution is estimated together with the fixed effects in the model. Notably, MML has a wide range of applications, including all types of IRT models, even multidimensional models. Additionally, the MML is highly efficient for parameter estimation, regardless of the length of the test. It also provides estimates for items with perfect score. Furthermore, MML data likelihoods can be used for model fit indices (Embreston & Reise, 2000, p.214). However, one limitation of MML is its computational complexity. It also usually requires an assumed distribution of person ability levels for item parameter estimation which may lead to biased estimates. Nevertheless, MML is currently the most popular estimation approach for IRT models.

When using MML for LG-EIRM estimation, the random effects are assumed to be normally distributed and the optimization of the marginal likelihood of the binary response data requires an integral that is intractable, which means there is no definite solution to the marginal likelihood maximization. To solve this problem, two types of

approximations are available in general, approximation to integral and approximation to integrand. Those approximating processes aim to provide an expression of the integral of the marginal likelihood optimization as a closed-form solution (DeBoeck & Wilson, 2004, p.347; Doran et al., 2007). Laplace's method (Tierney & Kadane, 1986) is a commonly applied approach that provides approximation to integrand. In Laplace's method, the integrand of the contribution of person p to the marginal likelihood is transformed to an exponent $\exp(\log(\Pr(y_p \mid \beta, \hat{\theta}_p)\phi(\theta_p \mid 0, \sigma_\theta{}^2)))$, where y_p is the response of person p, β is the item difficulty, θ_p is the latent trait of person p, and $\sigma_\theta{}^2$ is the variance of θ_p. This exponent is then approximated by a second-order Taylor series expansion about its maximum $\hat{\theta}_p$ (see Wilson & DeBoeck, 2004, p.352; Doran et al., 2007).

All ML estimation approaches can be implemented through different software programs, including IRT-specific software programs (e.g. jMetrik, IRTPRO, WINSTEPS, PARSCALE) and general statistical programs (e.g. SAS, R).

Early Literacy and Alphabet Knowledge

As the foundation of children's development of reading skills and strategies, early literacy skills acquired in kindergarten and early grades of school have attracted a lot of attention (Coyne & Harn, 2006). This section presents an overview of important components of early literacy and stresses on one of the core early literacy fundamentals, alphabet knowledge, which was the construct this study applies growth models to. In addition, particular letter-specific properties discussed in this section were used as item properties in the LG-EIRMs.

Early literacy mainly consists of four dimensions/components: phonological awareness, alphabet knowledge, concept of word, and grapheme-phoneme correspondence (Invernizzi et al., 2004). Phonological awareness is defined as the "awareness of sounds in spoken (not written) words that is revealed by such abilities as rhyming, matching initial consonants, and counting the number of phonemes in spoken words," (Stahl & Murray, 1994, p. 221). Alphabet knowledge refers to child's knowledge and ability to recognize all features of written letters, including letter forms, letter names, and letter sounds (Invernizzi, 2004; Huang, Tortorellu, & Invernizzi, 2014). Concept of word describes children's ability to "segment spoken sentences and phrases into words and to match spoken words with their counterparts text," (Invernizzi et al., 2004). Grapheme-phoneme correspondence refers to a child's ability to recognize the relationship between graphemes (i.e. letters) and corresponding phonemes (i.e. sounds) and to decode and write based on such relationship (Adams, 1990; Invernizzi et al., 2004). These four early literacy dimensions cover key abilities that can predict a child's future reading achievement. Therefore, Invernizzi, Sullivan, Meier, and Swank (2001, 2004) emphasized that in order to achieve early success in reading, children have to master those core constructs of early literacy. Among those core skills, alphabet knowledge is one of the constructs strongly associated with children's reading performance (Adams, 1990; NELP, 2008; Snow, et al., 1998). Scarborough (1998) conducted a comparison study which found alphabet knowledge could be as powerful a predictor as an entire early reading literacy test in terms of the ability to project children's future reading performance. Alphabet knowledge typically refers to a child's knowledge of letter names and letter sounds and the ability to recognize all aspects of letters

(Invernizzi et al., 2004; Puranik, Lonigan & Kim, 2011; Adams, 1990; Stevenson & Newman, 1986; Treiman, 2006). It is an essential element of early literacy that children need for reading and spelling (Adams, 1990; Piasta & Wagner, 2010; Lonigan et al., 2000). Assessment of alphabet knowledge usually includes tasks measuring children's ability to identify letter names and letter sounds (Piasta & Wagner, 2010; McBride-Chang, 1998; Treiman & Broderich, 1998; Levin & Ehri, 2009).

Among the two important components of alphabet knowledge, letter-name knowledge refers to children's ability to identify the names of letters given their corresponding graphic shapes (Huang and Invernizzi, 2014; Evans et al., 2006; Foulin, 2005). It makes significant contribution to children's visual recognition of words and acquisition of core literacy skills, especially spelling and reading (Foulin, 2005; McGee, Lomax, & Head, 1988; Adams, 1990). Huang and Invernizzi (2014) emphasized children who fail to obtain good letter-name knowledge and alphabet recognition skill will likely be at risk for future reading difficulties. However, Arciuli and Simpson (2011) pointed out all letters are not equally difficult because letter-specific features have impact on a child’s probability of naming a letter correctly (Evans et al., 2006). For example, visual confusability (i.e. shape confusability) is defined as the shape similarity of the letter to other letters (or numbers) (Huang & Invernizzi, 2014). Children may have greater probability of mistaking the letter with strong visual confusability for another letter (Ehri & Roberts, 2006; Treiman, 2006). This letter-specific property has been studied for its influence on children's performance in naming letters (Briggs & Hocevar, 1975; Cohn & Stricker, 1976; Evans et al., 2006; Fiset et al., 2008; Huang & Invernizzi, 2014). Considering this literature, the 26 alphabet letters can be classified as letters that are not

often confused (i.e. o, r, x), letter that are sometimes confused (i.e. a, c, e, f, s, t, y, z), letters that are often confused (i.e. i, j, k, l, m, w), and letters that are very often confused (i.e. b, d, g, h, n, p, q, u, v) (Huang & Invernizzi, 2014).

The other important component of alphabet knowledge, letter-sound knowledge, is defined as a child's ability to "provide the sounds associated with a particular letter form," (Huang et al., 2014). It plays an important role in a child's grasp of alphabetic principle, word decoding skills, and understanding of phonics instruction (Huang et al., 2014). As a result, children without a good mastery of letter-sound knowledge will very likely have difficulties in developing more complex reading and writing skills later (Hammill, 2004; Storch & Whitehurst, 2002). Huang and Invernizzi (2012) pointed out the difficulties of learning different letter sounds vary and some letter sounds are easier for children to learn than others, due to a set of factors associated with the characteristics of individual letters, such as the relative difficulty of identifying a letter's sound, the relationship between letter name and its sound, and the number of sounds the letter presents (Huang & Invernizzi, 2012; Evans et al., 2006; McBride-Chang, 1999). Particularly, the phonological relationship between letter sounds and their letter names (i.e. name-and-sound relationship) determines different letter-name structures, which affect how children use such name-and-sound relationship to learn letter sounds (Huang, et al., 2014; Evans et al., 2006). To be more specific, the letter-name structure property provides information about how letter sounds are related to their names and according to it, letters sounds can be classified into four categories, sounds associated with consonant-vowel (CV) pattern, sounds associated with vowel-consonant (VC) pattern, sounds not associated with letter's primary sound (NA) pattern, and vowel sounds (VO) (see Evans

et al., 2006; Treiman & Broderick, 1998; Huang et al., 2014). However, in regards of how letter-name structure affects relative easiness of letter sounds for children to learn are not entirely conclusive (Evans et al., 2006; McBride-Chang, 1999; Share, 2004; Treiman & Broderick, 1998; Huang et al., 2014).

The letter-associated properties have detectable impact on alphabet knowledge teaching and learning. Using the relationship between letter properties and alphabet knowledge instruction as rationales, various organizational patterns of alphabet knowledge instruction were formed and studied (Jones, Clark, & Reutzel, 2012; Rohrer & Pashler, 2010). Likewise, as to alphabet knowledge assessment, letter properties may have impact on examinee's responses to different test items since those letter properties are associated with children's learning advantages. Investigation of relationship among letter properties, test items, and examinee responses can help improve our understanding of children's developmental process of alphabet knowledge and inform the design of early literacy instruction and assessment.

Early Literacy Assessment

In recent years, growing attention has been focused on child's literacy development in kindergarten and even before (Missall & McConnell, 2010). It is believed that preschool years are a crucial period for educators to apply proper strategies to shape and positively impact children's literacy growth trajectories (VanDerHeyden, Snyder, Broussard, & Ramsdell, 2008). The beginning stage of kindergarten is especially critical to children to acquire reading and writing skills because their early literacy skills are in “a state of gradual maturation” (Invernizzi et al., 2004; Chaney, 1992; Snow, et al., 1998).

Those early literacy skills taught and assessed at kindergarten will precede and facilitate the later development of conventional reading skills of children (Invernizzi, 2004). Additionally, there is an increasing interest on child-specific information and its connection to his/her early literacy developmental progress because such progress is strongly associated with the child's later academic success (Missall & McConnell, 2010).

Early literacy assessments are a vital part of understanding a child's early literacy development. The National Association for the Education of Young Children (NAEYC, 1991) described the key purposes of early literacy assessment as "to plan instruction for individuals and groups and for communicating with parents" and "to identify children who may be in need of specialized services or intervention." As suggested by NAEYC (1991), one of the main purposes of early literacy assessment is to help educators identify children with risk of future reading difficulty at early stage of schooling. Notably, Dickson and Neuman (2007) argued that the emphasis should be placed on the identification and planning of instructional support to those children at risk rather than sole classification of children in the testing pool. Compared to later remediation, early intervention is a means with less cost and higher efficacy (Heckman & Masterov, 2007). Therefore, the screening results are used to initiate appropriate early interventions for those children with problems acquiring early literacy skills. Justice, Invernizzi, and Meier (2002) also noted that assessment outcomes can be used for timely detection of difficulties of children in literacy achievement and for guidance of intervention design and implementation. For instance, Helman (2005) suggested interpretation of results from early literacy assessments may provide useful information for instructing non English Language Learners (ELL) to learn core early literacy skills. Speech-language pathologists

can also use information extracted from early literacy assessment to help design their instruction or intervention (Justice et al., 2002).

To address the functions of early literacy assessment more specifically, Invernizzi et al. (2004) summarized that an appropriate early literacy assessment tool should be able to serve four purposes: screening, diagnosis, progress monitoring, and outcome assessment. She pointed out a successful systematic assessment should be able to provide functions to document children's early literacy achievement as well as to link achievement information to instruction planning. In practice, most assessments function as instruments to screen children for school readiness, identify children with reading difficulty, hold teacher or school accountable for children's achievement or funding expenditure, and inform the design of specific instruction (Dickinson& Neuman, 2007). To illustrate how those functions are implemented, Coyne and Harn (2006) provided a series of school-based empirical examples. In Coyne's study, the DIBELS assessment was used to measure early literacy skills and inform school on instruction-relevant decision making. First, the assessment tool demonstrates its power of screening by identifying the group of early-grade children at risk of developing reading skills when the fall semester starts. Based on the outcome, the research-based reading instructional plan and intervention program are implemented. As for the progress monitoring function, the results from weekly literacy assessment help teachers decide if a group of children at risk for reading problems are making adequate growth towards standards. Additional instructional adjustments could be prepared based on those assessment data. As showed by Coyne's examples, the data from multiple early literacy assessments allow schools and teachers to access important information about children's initial performance level and

their growth trajectories. Therefore, the validity and perceived importance of instructional efforts and intervention programs targeting those children with reading difficulties would increase (Coyne & Harn, 2006).

From the perspective of children who experience difficulty in understanding and acquiring those literacy skills, benefits of having early literacy assessment are significant. The result of such assessment helps educators accurately identify domains or areas of needs and craft assistance to those children to neutralize the impact of any problem on later reading ability attainment (Snow et al., 1998; Invernizzi et al., 2004). For lots of states, the ultimate goal is to develop and implement effective and efficient strategies to prevent reading difficulty of children and enhance the efficacy of classroom instruction. The success of achieving such goal depends on precise identification of children who have requests of customized intervention and continuous monitoring of their literacy development progress (Invernizzi et al., 2004).

In recent decades, a variety of assessments have been developed and widely applied to measure children's performance of early literacy skills, such as Phonological Awareness Literacy Screening (PALS), Dynamic Indicators of Basic Early Literacy Skills (DIBELS), and AIMSweb, etc. (Invernizzi et al., 2011; Deno & Fuchs, 1987; Pearson, 2012a). The availability of these assessments allows for not only the monitoring of children's developmental growth and but also the examination of relationship among children's growth, test characteristics, and children properties.

For any early literacy assessment instrument, the utmost important key to successful and effective measurement of the construct of early literacy skills is sound psychometric quality. More specifically, the reliability and validity of a measure reflect

its psychometric soundness which directly affect the degree of to which we can trust the score inferences. Reliability is generally defined as an index indicating the degree of consistency between two sets of test scores produced from two similar hypothetical or practical measurement processes (Meyer, 2010). Validity refers to "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (Messick, 1990, p.1). As suggested, validity is a matter of degree, not all or none.

For example, DIBELS is a measurement instrument for assessing essential early literacy and reading skills from kindergarten to 6th grade with the purpose to identify children who are at risk acquiring basic early literacy skills and need early additional support to prevent them from experiencing later reading difficulties. It comprises a set of measures which respectively assess phonemic awareness, phonics, accuracy and fluency with connected text, vocabulary and language skills, and reading comprehension. Good et al. (2011) addressed technical properties about DIBELS which showed it has consistently high reliability coefficients across all skill domains within the construct being measured. Strong content validity, criterion-related validity, and discriminant validity evidence of DIBELS were also reported, which suggests it is an effective and trustworthy instrument. Their report also emphasized DIBELS's sensitivity to children's developmental growth in early literacy in response to interventional support in different skill domains. National Reading Panel (2000) described DIBELS measures as key indicators of foundational early reading. Another example is AIMSweb, a web-based assessment for early literacy and early numeracy skills for Grades K-12. It provides schools and teachers with

functions of universal screening, progressing monitoring, and data management (Pearson, 2012a). One of AIMSweb measures, test of early literacy, contains tasks of letter naming fluency, letter sound fluency, phonemic segmentation fluency, and nonsense word fluency. The reliability of this measure was addressed from three aspects, the alternate-form reliability with average value above 0.80, and the test–retest reliability above 0.85, and the inter-rater reliability above 0.82. Its criterion validity was demonstrated by decent correlations with other similar assessments (i.e. Woodcock-Johnson revised broad reading, Woodcock-Johnson revised reading skills, test of phonological awareness, teacher rating, and developmental skills checklist) which ranges between 0.44 and 0.75 on average (Pearson, 2012a).

The PALS is an assessment tool that measures a child's knowledge of early literacy fundamentals that are effective predictors of future reading performance. The major purpose of PALS is to identify children who have difficulty reaching certain performance standards and may need additional intervention (Invernizzi, Juel, Swank, & Meier, 2011). PALS-K is the version of PALS administered to kindergarteners.

The conceptual framework of PALS-K (Invernizzi et al., 2011) involves two components, phonological awareness and literacy skills, that are measured through six required subtests and one optional subtest (see Table 2.1).

Table 2.1

Conceptual Framework for PALS-K

Component	Subtest
Phonological Awareness	Rhyme Awareness Beginning Sound Awareness
Literacy Skills	Alphabet Knowledge Letter Sounds Spelling Concept of Word Word Recognition in Isolation (optional)

Invernizzi et al. (2011) report that PALS-K has test-retest reliability estimates that range from 0.78 to 0.95, internal consistency estimates from each subtest that average 0.86, and inter-rater reliability estimates that range between 0.96 and 0.99. Validity evidence for PALS-K consists of content validity, criterion-related validity, and construct validity. Its content validity is ensured by careful item selection by experts. Criterion-related validity is addressed by comparing PALS-K scores to Stanford Achievement Test (1996) scores. The results of comparison showed there were medium to high correlations. Finally, construct validity has been assessed by evaluating the internal structure of PALS-K and classical item analysis and results indicated its internal consistency.

Huang and Konold (2014) studied the structure of PALS-K and their results indicated the actual factorial structure of PALS-K data is slightly different from its original conceptual framework. The result from their confirmatory factor analysis suggested the best-fit factorial structure model for PALS-K is a three-level hierarchical model, where a single second-order factor, early literacy, has influence on three first-

order factors, phonological awareness, alphabet knowledge, and contextual knowledge. The model presented by Huang and Konold (2014) is shown in Figure 2.2 (error terms are not shown in the graph). In this model, two subtests are constructed to measure each first-order factor. Aligned with PALS-K conceptual framework, there are six required subtests in total. Among the three first-order factors, alphabet knowledge consists of the alphabet recognition subtest and the letter sounds subtest. The letter sounds subtest will be where the interest of this study is centered on.

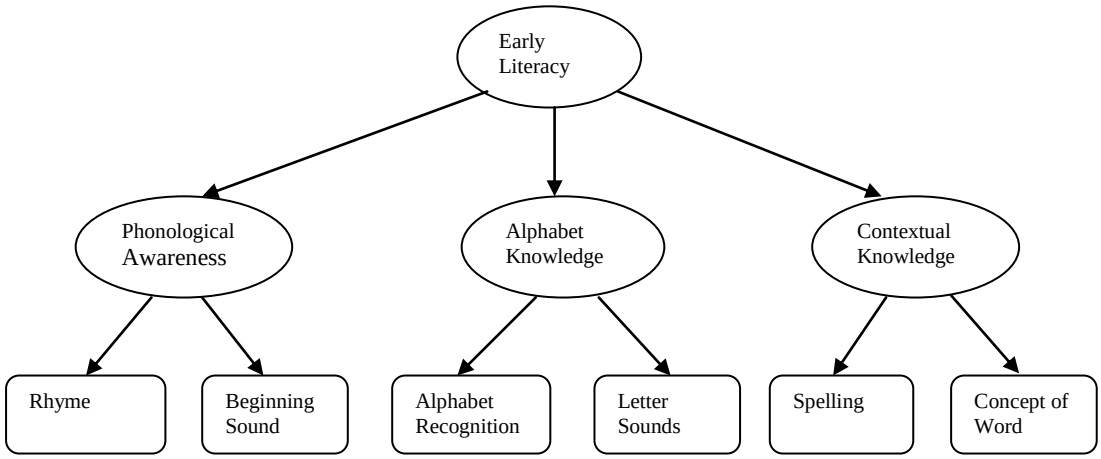

Figure 2.2 Factor Structure of PALS-K Data

Early Literacy Development Studies

Access to early literacy assessment data has significant practical meanings to tracking a child's learning progress, and as a result, research interest on early literacy development and growth has been rising. One benefit from early literacy growth studies is to obtain information about the relation between early reading performance and later reading achievement (Dickinson, Tabors, & Roach, 1996). Another benefit of analyzing growth data obtained from early literacy assessments is to explore and understand the

influences of different factors on literacy development of children. The information about sources of individual differences in growth can guide customization of instructional support provided by teacher or school.

Some early literacy longitudinal studies examined the predictive validity of important early literacy skills for later literacy performance, instead of estimating actual growth in either observed score or latent construct. The data used in those studies are usually collected across multiple years. Lonigan, Burgess, and Anthony (2000) conducted a study to examine the predictive significance of some emergent literacy skills of preschool children. Participants were two groups of preschool children whose performance were tracked from early to late preschool and from late preschool to kindergarten or first grade respectively. They used SEM to build longitudinal latent variable models with those literacy tasks as latent variables and subtask scale scores as observed variables. Their models merely focus on examining the relations between emergent literacy skill factors and later literacy and reading skill factors. Results of their study revealed phonological sensitivity and letter knowledge demonstrated strong predictive relations with later reading abilities and significant contributions to the influence on children's literacy development. A similar study conducted by McCormick and Haack (2010) investigated the predictive validity of Early Literacy Individual Growth and Development indicators (EI-IGDIs; Missall & McConnell, 2010) for child's later academic success. It was a longitudinal study that examined the relations between children's literacy skills in prekindergarten measured by EI-IGDIs and later literacy skills in kindergarten through 2nd grade measured by DIBELS and Measures of Academic Progress (MAP; NWEA, 2011). Multiple regression was utilized to determine if the EI-

IGDIs score and DIBELS score were significant predictors of MAP score. Their study suggested the successful instructional approach and school district's appropriate education strategy could help improve children's growth in early literacy and positively influence their later literacy success.

Another group of early literacy growth studies focus on examining children's literacy and reading development based on their absolute growth in either observed variables or latent variable. Those studies collected multiple-wave data within one year and used growth curve modeling for data analysis, which permitted the study to track growth trajectory and analyze growth trend (Pan, Rowe, Singer, & Snow, 2005; Speece, Ritchey, Cooper, Roth, & Schatschneider, 2004; McCoach,et al. 2006; Hammer, Lawrence, & Miccio, 2007). For example, McCoach,et al. (2006) used a multilevel growth curve model to estimate initial status and growth rate of children's performance of Early Childhood Longitudinal Study—Kindergarten cohort (ECLS–K; Tourangeau et al., 2009) over first 2 years of school. ECLS–K measures a set of core early literacy skills. Four waves of longitudinal data of children's performance were collected from kindergarten to 1st grade. Additionally, child property variables (e.g. socioeconomic status, ethnicity) and school-related (e.g. percentage of minority students, percentage of free-lunch students) variables were added to the model to examine their contribution to between-school and within-school variability. Results suggested overall, children grow much faster in reading skills in 1st grade than what they did in kindergarten. As for variances in initial performance and growth rate of children among schools, those child property variables explained a significant portion of it. Another example is the growth study conducted by Gutierrez & Vanderwood (2013). Their study investigated early

literacy growth of 2nd graders and examined the variations among ELLs at different English language proficiency levels. Children's skills in phonological awareness, alphabetic principle, and oral reading fluency were measured in the fall, winter, and spring of 2nd grade and three-wave data were collected. Assuming a linear growth, this study used a 2-level growth curve model to analyze the individual differences in growth trajectory of the observed scale scores. Their result revealed the initial status and growth trajectory were significantly different among different English-proficiency groups of examinees. In addition, children's growth in different components of early literacy varied. One notable feature of this type of growth studies is they offered the opportunity to examine concurrent development of children in those early literacy skills since measurement data were available at the beginning and the end of a school year.

Overall, the above studies were interested in investigating growth in some important early literacy skills which were addressed earlier in this chapter, including phonological awareness, alphabet knowledge, word decoding, etc. They were also interested in examining the impact from person property covariates (such as English language learner, gender, age, socioeconomic status) on differences in growth trajectories among children. However, one big limitation of those studies is most of them were not using latent variable growth models to measure the latent growth. Also, even the SEM approach was used to construct latent variable models, they used scale score or item parcel scores, instead of categorical item-level response data, assuming that scores are normally distributed. They did not take characteristics of individual items into consideration even those characteristics may have influence on children's growth. Although few studies on alphabet knowledge discussed earlier (Huang and Invernizzi,

2012; Huang et al., 2014) used item-level response as dependent variables in their logistic regression models, only cross-sectional data, rather than longitudinal data, were analyzed. To address those limitations, my study utilized IRT-based latent growth models and item-level response data to examine child's developmental growth in alphabet knowledge within one school year. Item properties (e.g. letter-shape confusability, letter-name structure) and important examinee properties (e.g. disability status, age, ELL status) were also examined with respect to their relationship with latent growth.

Research Questions

Due to the importance of understanding a child's developmental progress in alphabet knowledge, I utilized a series of LG-EIRMs to investigate the latent growth of alphabet knowledge including alphabet recognition and letter sounds within one school year. Additionally, multiple factors associated with individual differences in growth were examined. Moreover, the lack of time-invariance of item properties was tested. Therefore, this study provided answers to the following questions:

1. What is the amount of average latent growth in alphabet recognition and letter sounds respectively over three time points?

2. What examinee properties (e.g. pre-k schooling, English Language Learner status) influence the latent growth?

3. Does any item property (i.e. letter-shape confusability and letter-name structure) show a lack of time-invariance (i.e. drift)?

Chapter 3 Method

PALS-K Measure

The PALS is an assessment for children that measures knowledge of early literacy fundamentals that are effective predictors of future reading performance. A major purpose of PALS is to identify children who have difficulty reaching certain performance standards and may need additional intervention (Invernizzi, Juel, Swank, & Meier, 2011). PALS-K is the version of PALS administered to kindergarteners. PALS-K has two primary forms in use, Form A and Form B, which are designed to be parallel and are given to the examinees in alternating school years. The same test form is used for a whole school year and the test is administered once in fall and once in spring. An optional test form, Form C, is administered in midyear.

PALS-K has exhibited sound psychometric quality and technical adequacy in prior research. Invernizzi et al. (2011) report that PALS-K has test-retest reliability estimates that range from 0.78 to 0.95, internal consistency estimates from each subtest that average 0.86, and inter-rater reliability estimates that range between 0.96 and 0.99. Validity evidence for PALS-K consists of content-, criterion-, and construct-related forms of validity evidence. Their report has also evaluated PALS-K items for differential item functioning (DIF) using the Mantel-Haenszel (MH) statistic to compare children in need of additional instruction to those who do not need extra instruction. The result revealed no substantial amounts of DIF between the two groups.

Organization of PALS-K Measure

The conceptual framework of PALS-K (Invernizzi et al., 2011) includes two main components, phonological awareness and literacy skills, which are measured by six required subtests and one optional subtest together (see Table 2.1). However, according to Huang and Konold (2014)'s study, empirical data indicated the actual factorial structure of PALS-K data is a three-level hierarchical model, where a single second-order factor, early literacy, has influence on three first-order factors which can also be seen as three constructs: phonological awareness, alphabet knowledge, and contextual knowledge (see Figure 2.2). In their model, each first-order factor/construct is measured by two subtests. According to PALS-K conceptual framework, there are six required subtests in total which are the Rhyme Awareness task, Beginning Sound Awareness task, Alphabet knowledge task, Letter Sounds task, Spelling task, and Concept of Word task. These six subtests are aggregated to create the examinee's sum score.

Among the required subtests, the alphabet knowledge (also named lower-case alphabet recognition) task and Letter Sounds task are used to measure a child’s performance on alphabet knowledge, a construct that refers to knowledge of letter names and letter sounds and the ability to recognize all aspects of letters (Invernizzi, 2004; Puranik, Lonigan & Kim, 2011; Adams, 1990; Stevenson & Newman, 1986; Treiman, 2006).

Letter-name knowledge is usually described as a child’s knowledge of the names of all of the letters of the alphabet in upper and lower case. It makes significant

contribution to a child's visual recognition of words and acquisition of core literacy skills, especially spelling and reading (Foulin, 2005; McGee, Lomax, & Head, 1988; Adams, 1990). However, Arciuli and Simpson (2011) pointed out all letters are not equally difficult because letter-specific properties impact a child's probability of naming a letter correctly (Evans et al., 2006). Among those properties, visual confusability (i.e. shape confusability) refers to the shape similarity of the letter to other letters (or numbers) (Huang & Invernizzi, 2014). Children may have a greater probability of mistaking the letter with strong visual confusability for another letter (Ehri & Roberts, 2006; Treiman, 2006).

In PALS-K, the lower-case alphabet recognition task measures a child's ability to provide names of all lower-case alphabet letters. It includes 26 items representing 26 lower-case alphabet letters which can be classified into groups based on letter-specific shape confusability property. All items are binary items with an incorrect answer scored as 0 points and a correct answer scored as 1 point. Table 3.1 shows item property groups of the 26 alphabet recognition items of PALS-K based on shape confusability:

Table 3.1

Item Properties of Lower-case Alphabet Recognition Subtest

Shape confusability	Lower-case alphabet recognition items
Not often confused (NOFC; shape1)	o, r, x
Sometimes confused (SC; shape2)	a, c, e, f, s, t, y, z
Often confused (OFC; shape3)	i, j, k, l, m, w
very often confused (VOFC; shape4)	b, d, g, h, n, p, q, u, v

Not only does the shape of a letter affect a child's ability to learn it, but also the sound of a letter. Huang and Invernizzi (2012) pointed out that the difficulty of learning different letter sounds varies and some letter sounds are easier for children to learn than others. Particularly, the letter-name structure provides information about how letter sounds are related to their letter names, and different letter-name structures have effect on how children utilize name-and-sound relationship to learn letter sounds (Huang, Tortorellu, & Invernizzi, 2014). Letters can be classified into four groups: sounds associated with letter names in a consonant-vowel (CV) pattern (i.e. b, d, j, k, p, t, v, z), sounds associated with letter names in a vowel-consonant (VC) pattern (i.e. s, r, f, l, n, m), sounds associated with letter names unrelated to their primary sounds (NA; i.e. w, h, c, y, g), and sounds associated with vowel-sound letters (VO; i.e. a, e, i, o, u; see Evans et al., 2006; Treiman et al., 1998; Huang, Tortorellu, & Invernizzi, 2014).

The Letter Sounds task of PALS-K measures a child's ability to produce sounds associated with individual letters. It includes 26 binary items with an incorrect answer scored as 0 points and a correct answer scored as 1 point. Table 3.2 shows the item property groups of the 26 letter-sound items of PALS-K based on letter-name structure property:

Table 3.2

Item Properties of Letter Sounds Subtest

Letter-name structure	Letter-sound items
CV (sound1)	B,T, J, K, V, P, Z, D
VC (sound2)	S, R, F, L, N
VO (sound3)	O, A, I, U, E
NA (sound4)	W, H, C, Y, G
Digraph (sound5)	Ch, Sh, Th

Notably, the alphabet recognition subtest consists of 26 lower-case alphabet letters as 26 individual items and the letter sounds subtest contains 23 upper-case alphabet letters and 3 digraphs together as 26 individual items. Therefore, 23 letters from the alphabet string appear on both subtests but in different formats, lower-case and upper-case respectively. Different formats of the same letter are treated as two separate items belonging to different subtests. For example, the lower-case letter “a” appeared as one item of the alphabet recognition subtest, and the upper-case letter “A” is an item in the other subtest, letter sounds.

As described below, growth analysis was conducted on the alphabet recognition subtest and the letter sounds subtest separately. Models for each subtest include coefficients for the item properties in Tables 3.1 and 3.2 in addition to other coefficients that represent person properties.

Sample

This study explored the magnitude of growth in the construct of Alphabet Knowledge among kindergarten children. The examinee sample used in the study included 5,000 kindergarten children from a mid-Atlantic state. Data comprised item-level responses to the Lower-case Alphabet Recognition subtest and Letter Sounds subtest from PALS-K that measured during fall, mid-year, and spring of the 2013-2014 school year.

Although Form A was used for fall and spring test administrations and Form C was used for mid-year test administration, items of both the alphabet recognition subtest and the letter sounds subtest stay constant across test forms. Responses to total 52 alphabet knowledge items were extracted from the PALS-K database, along with demographic variables of interest, including pre-k schooling , disability status, ELL status, and age in fall. The 5,000 examinees used in the study were randomly selected from a larger sample of examinees who took PALS-K during the 2013-2014 school year. Table 3.3 gives descriptive statistics of the 5,000 sample.

Table 3.3

Descriptive Statistics of the Sample

Variable		Number	% of total
Gender	Male	2,572	51.44%
	Female	2,428	48.56%
Pre-k schooling	No	2,275	45.50%
	yes	2,725	54.50%
ELL status	non-ELL	4,699	93.98%
	ELL	301	6.02%
Disability status	No	4,680	93.60%
	yes	320	6.40%
Age in fall	Mean= 66.80 months		
	SD=4.09 months		

Person properties, including pre-k schooling, ELL status, disability status, and age in fall, were included as person properties in some growth models described below to understand their effect on child's latent ability and growth in alphabet knowledge. All person property variables are categorical except for a child's age in fall 2013 which is a continuous variable. Table 3.4 lists all the person properties in this study and how they were coded in the dataset.

Table 3.4

Person Properties

Person property	Person groups
ELL status	non-ELL (ell=0) ELL (ell=1)
Pre-k schooling	non-pre-k(prek=0) pre-k (prek=1)
Disability status	non-disability (dis=0) disability (dis=1)
Age in fall	Mean=66.80 months Standard Deviation=4.09 moths

Latent Growth Models

EIRMs use an explanatory approach to reveal the nature of the relationships among item response and item- or person-properties in the model. Recently, EIRMs have been extended to latent growth studies. Several LG-EIRMs have been developed and tested using empirical data, including Embretson(1991)'s MRMCL, Wilson et al.'s (2011) LG-IRM, Pastor and Beretvas (2006)'s P-HGLM, and Cho et al.'s (2013) GELIRM. Similar to regular EIRMs, LG-EIRMs aim to provide explanation of item responses through effects contributed by person properties, item properties, or both, and they also aim to help improve the understanding of what certain person or item characteristics influence those responses.

As a latent growth item response model, Embretson (1991)'s MRMLC, was used as the framework for this study to build a series of LG-EIRMs. As shown in equation 2.9, for an item given at $t = 1,...,T$ time points, the MRMLC is given by

$\eta_{pit} = \sum_{t=1}^{T} \theta_t Z_{pt} + \varepsilon_p - \beta_i$. Extending it to include additional person and item property covariates, equation 2.10 presents the LG-EIRM by incorporating RW-LLTM elements, which is $\eta_{pit} = \left(\sum_{j=1}^{J} \theta_{jt} Z_{pjt} + \varepsilon_p \right) - \left(\sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \right)$. Importantly, to test the invariance of item parameters or item properties while analyzing growth, the LG-EIRM shown above can be revised to incorporate a time-by-item interaction. It can also be extended to include person-by-item interaction terms to study the effect of group membership on latent growth.

This study used a series of LG-EIRMs to study growth in alphabet recognition and letter sounds and test for item property time-invariance. Given research about item properties in measures of early literacy and what is known about the development of alphabet knowledge, this approach is well-suited for alphabet knowledge measures. Equation 3.1 (also see equation 2.12) shows the general form of LG-EIRM for unidimensional Rasch model-based data with interaction term to test time-invariance of items:

$$\eta_{pit} = \left(\sum_{j=1}^{J} \theta_{jt} Z_{pjt} + \varepsilon_p \right) - \left(\sum_{k=1}^{K} \left(\beta_k X_{ik} + \gamma_k V_{tik} \right) + \varepsilon_i \right) \qquad (3.1)$$

where the parameter $\theta_{\bullet 1}$ reflects the change in latent trait from time one to time two, and $\theta_{\bullet 2}$ indicates the amount of change from time two to time three, Z_p is a matrix coded with a Weiner simplex process (Embretson,1991), ε_p indicates variance of individual examinee's initial status and growth on the latent trait after group effects are accounted

for which is normally distributed with zero means and variances $\sigma^2_{\varepsilon_p}$. X_{ik} is equals 1 if item i involves item property k $(k = 1,...,K)$ and zero otherwise, β_k is the regression weight of item property k, γ_k is the time-by-item interaction effect, V_{tik} is the indicator of interaction effect which equals to the product of the time-point indicator and an item indicator, and ε_i represents the error term of item i which is normally distributed with zero means and variances $\sigma^2_{\varepsilon_i}$.

Data Analysis Procedures

Data analysis included two major steps: psychometric evaluation of data from each single time point and latent growth modeling of three-time-point longitudinal data.

Step one. Descriptive analysis and classical item analysis were performed on item-level response data of the sample. Descriptive statistics of the sum score of the alphabet-recognition subtest and the letter-sounds subtest at each test occasion provided basic statistical characteristics of the distribution of each subtest sum score at each time point. Classical item analysis provided classical item difficulties and discriminations of all items. The classical reliability of each subtest was also evaluated respectively as an indication of the general psychometric quality of the subtest.

Step two. Multiple LG-EIRMs were constructed and applied to the three-time-point item response data of alphabet recognition subtest and letter sounds subtest respectively. As the two components of the alphabet knowledge construct, alphabet recognition ability (i.e. letter-name knowledge) and letter-sound knowledge are each treated as unidimensional construct being measured by their corresponding subtest. The

unidimensionality of the item response data to each subtest was therefore assumed. As mentioned earlier, the unidimensional Rasch model-based LG-EIRM (see Equation 2.9) was utilized as the base model and multiple LG-EIRMs were built upon it and compared in the current study.

To answer all research questions, modifications were made on the most general model (see Equation 3.1) to construct the following types of LG-EIRMs:

1. *Item properties model* (M1). This model is the most basic model to be tested. Estimates from this model provide information about amount of latent growth across time points and mean difficulty of each item property (see Tables 3.1 and 3.2). This model answers research question 1 regarding the latent growth over three time points, when neither lack of time-invariance of item properties nor person group differences in growth is taken into consideration. It is given by Equation 3.2 which includes fixed effects of average growth (i.e. time effect) of person p , fixed effects of item properties, random effect of individual initial status and growth of person p , and random error term for item i :

$$\eta_{pit} = \left(\theta_t Z_{pt} + \varepsilon_p\right) - \left(\sum_{k=1}^{k} \beta_k X_{ik} + \varepsilon_i\right) \qquad (3.2)$$

2. *Item properties and time-by-item property interaction models* (M2 – M5/M2 – M6). The main purpose of these models is to answer research question 3 and determine whether any item property exhibits a lack of time-invariance (i.e. drift). If these models do not fit better than M1, then time invariance of item properties may be assumed and item property by time interactions will be omitted from subsequent models. On the other hand, if these models fit better than M1, then subsequent models must include the

interactions to account for drift. These models shown by Equation 3.3 include fixed effects for average growth, item properties, and the interaction of time by an item property (i.e. drift). Random effects include individual initial status and growth of person p at times two and three, and a random error term for item i :

$$\eta_{pit} = \left(\theta_t Z_{pt} + \varepsilon_p\right) - \left(\sum_{k=1}^{K}\left(\beta_k X_{ik} + \gamma_k V_{tik}\right) + \varepsilon_i\right) \tag{3.3}$$

3. *Item properties and person properties model* (M7/M8). Estimates from this model provide information about average latent growth across time points, difficulties of item properties, and group differences in latent ability. It is given by Equation 3.4, which includes fixed effects for average growth, item properties, and overall difference in latent ability between examinees who belong to person group j and examinees who do not, and random effects for individual initial status and growth of person p and an error term for item i . It is given by Equation 3.5 when controlling for a lack of time-invariance of item properties. Two equations are presented as:

$$\eta_{pit} = \left(\sum_{j=1}^{J}\theta_{jt} Z_{pjt} + \varepsilon_p\right) - \left(\sum_{k=1}^{k}\beta_k X_{ik} + \varepsilon_i\right) \tag{3.4}$$

$$\eta_{pit} = \left(\sum_{j=1}^{J}\theta_{jt} Z_{pjt} + \varepsilon_p\right) - \left(\sum_{k=1}^{K}\left(\beta_k X_{ik} + \gamma_k V_{tik}\right) + \varepsilon_i\right) \tag{3.5}$$

In Equation 3.5, γ_k is the time-by-item property interaction effect and V_{tik} is the indicator of interaction effect which equals to the product of the time-point indicator and an item property indicator X_k . If significant drift is identified, Equation 3.5 will be used instead of Equation 3.4.

4. *Item properties, person properties, and person property-by-time interaction model* (M8/M9). Estimates of interest from this model include group differences in latent growth and difficulties of item properties. This model is able to answer research question 2 regarding the impact of examinee properties on latent growth. It is given by Equation 3.6, which includes fixed effects for average growth, item properties, overall difference in latent ability between different person groups categorized by person property j, and average difference in growth between person groups categorized by person property j. Random effects include individual initial status and growth of person p and an error term for item i. M8/M9 is given by Equation 3.7 if accounting for a lack of time-invariance of item properties. Two equations are presented as:

$$\eta_{pit} = \left(\sum_{j=1}^{J} (\theta_{jt} Z_{pjt} + \delta_j W_{pjt}) + \varepsilon_p \right) - (\sum_{k=1}^{k} \beta_k X_{ik} + \varepsilon_i) \tag{3.6}$$

$$\eta_{pit} = \left(\sum_{j=1}^{J} (\theta_{jt} Z_{pjt} + \delta_j W_{pjt}) + \varepsilon_p \right) - \left(\sum_{k=1}^{K} \left(\beta_k X_{ik} + \gamma_k V_{tik} \right) + \varepsilon_i \right) \tag{3.7}$$

where δ_j is the time-by-person property interaction effect, W_{pjt} is the indicator of interaction effect which equals to the product of the time-point indicator and a person property indicator Z_j. Notably, M8/M9 is the most complex and most general model to be tested in this study.

To obtain LG-EIRM parameter estimates, the Laplace approximation method was used for most models as implemented in lme4 package (Bates, Mächler, Bolker, & Walker, 2014) in R. Due to the large size of data and the complexity of some models, a

Bayesian-based MML method was used to incorporate priors on fixed effects and the random effect covariance matrix. This change to the estimation allowed the complex LG-EIRMs converge and produce parameter estimates. Bayes modal estimation methods were implemented using the blme package (Dorie, 2015) in R.

Model comparison. All LG-EIRMs are compared based on multiple model information criteria, including log likelihood ratio test (Wilks, 1938; Mood & Graybill, 1963), Akaike Information Criterion (AIC; Akaike, 1998), and Bayesian Information Criterion (BIC; Schwarz, 1978).

Chi-square test of log likelihood ratio (LR test) is the most commonly used model comparison method. It basically includes two steps: fitting models to the data using a maximum likelihood criterion and performing a chi-square test to compare models based on the log likelihood ratio statistics (Wilks, 1938; Mood & Graybill, 1963; Busemeyer & Wang, 2000). The LR test technically measures the change of likelihood discrepancy from one model to another model. However, its main limitation is that it is only applicable to comparisons among nested models. Another limitation is its dependency on sample size when picking better models. For example, it tends to pick the over-complex model when sample size is large and statistical power is high (see Cudeck & Browne, 1983; Busemeyer & Wang, 2000).

To overcome the limitations of LR test, AIC and BIC became widely applied model selection criteria because they permit comparisons of both nested and non-nested models that may differ in the number of free parameters. The model that minimizes AIC

and BIC values should be selected as the better-fitting model (Busemeyer & Wang, 2000). AIC and BIC can be expressed mathematically in Equation 3.8 and 3.9:

$$AIC = -2 \times ln(L) + 2 \times k \qquad (3.8)$$

$$BIC = -2 \times ln(L) + (\ln N) \times k \qquad (3.9)$$

where L is the maximized value of the likelihood function of the model, k is the number of model parameters to be estimated, and N is the number of individual cases in the sample. As shown above, they differ only by how strongly they panelize the large model. Normally speaking, using BIC tend to select models that are more parsimonious than the models chosen by AIC (Kadane & Lazar, 2004).

Through comparisons, the best-fitting LG-EIRM was selected and its estimates were particularly discussed.

Chapter 4 Results

This chapter presents the results from data analyses, including descriptive statistics, results of the psychometric property analysis of item response data at each time point, and results from latent growth modeling.

Descriptive Statistics of Sum Scores

Data includes item responses to 26 alphabet recognition (ABC) items and 26 letter sounds (LS) items from PALS-K. Data were collected in fall, mid-year, and spring of the 2013-2014 school year.

The average sum score of each subtest increased at each time point, while the standard deviation decreased over time (see Table 4.1). In addition, sum scores became increasingly negatively skewed and leptokurtic on each occasion. These statistics indicate that for both subtests, scores became higher and less variable over time. Moreover, the amount of growth resulted in a ceiling effect with average scores near the maximum possible score at the last time point.

Table 4.1

Descriptive Statistics of Subtest Sum Scores

		Mean	SD	Min	Max	Skewness	Kurtosis
Alphabet Recognition (ABC)	Fall	19.28	7.57	0	26	-1.09	2.92
	Mid-year	23.98	3.91	0	26	-3.20	14.76
	Spring	25.30	2.15	0	26	-6.30	53.10
Letter Sounds (LS)	Fall	13.57	7.76	0	26	-0.30	1.85
	Mid-year	20.73	5.35	0	26	-1.64	5.86
	Spring	24.10	3.35	0	26	-3.58	20.26

Classical Item Analyses

To evaluate basic psychometric properties of each item, I performed a classical item analyses using jMetrik (Meyer, 2014) on ABC and LS subtests separately. Results suggested difficulties and discriminations of all 52 items were within the appropriate range. No item exhibited abnormality, such as irregular value of difficulty or negative value of discrimination. Thus, we conclude that all items maintained proper basic psychometric characteristics across three test administrations. Descriptive statistics of classical item difficulty and discrimination and reliability coefficients of the two subtests are reported in Table 4.2:

Table 4.2

Classical Item Analysis Results

		Difficulty			Discrimination			Reliability	SEM
		Mean	SD	Range	Mean	SD	Range		
ABC	Fall	0.74	0.11	0.48~0.94	0.66	0.08	0.42~0.74	0.96	1.60
	Mid-year	0.92	0.06	0.76~0.99	0.56	0.09	0.34~0.67	0.92	1.11
	Spring	0.97	0.03	0.89~0.997	0.52	0.10	0.30~0.65	0.89	0.71
LS	Fall	0.52	0.22	0.10~0.87	0.62	0.12	0.34~0.73	0.95	1.76
	Mid-year	0.80	0.19	0.34~0.98	0.55	0.07	0.42~0.65	0.92	1.52
	Spring	0.93	0.08	0.69~0.99	0.53	0.04	0.45~0.61	0.89	1.10

As shown by Table 4.2, average item difficulty for both subtests rose over time, which means the average number of students who answered the items correctly increased over the three test occasions of the year (i.e. items became easier). Item difficulty also became less variable over three time points as indicted by decreasing standard deviations at each time point. The discriminations for all items across all time points ranged from 0.30 to 0.74 which indicated good discriminating ability of items Item discrimination decreased slightly over time, but item discrimination for ABC was slightly higher than item discrimination for LS at each time point.

The reliability of the ABC subtest ranged from 0.89 to 0.96 and the reliability of the LS subtest ranged from 0.90 to 0.95 over three test occasions. For both subtests reliability slightly decreased over time, but this result is most likely due to the decreasing variance in sum scores (i.e. more homogenous scores). Taken together, results indicate both subtests at each time point have good internal consistency and sound psychometric properties.

LG-EIRM Analysis

LG-EIRM growth analyses were performed on ABC subtest and LS subtest separately and results for each subtest are presented in separate parts of this section.

Lower-case Alphabet Recognition (ABC). As shown by ABC-M1 in Table 4.3, average latent growth was 2.33 at mid-year and 4.08 in spring on the logit scale, both of which were statistically significant. Random effect estimates shown by ABC-M1 in Table 4.4 indicate the variance of examinee initial ability was 7.07, 2.14, and 1.65 in fall, mid-year, and spring, respectively. Individual differences in ABC were large initially, but considerably smaller by the end of the year. Thus, not only did students improve learning across time points, but their performance became less variable. The fixed and random effects also support the notion of a ceiling effect such that by the end of the year most students can recognize all of the letter names. Looking at the correlations between person estimates (see Table 4.4), initial latent ability in fall is negatively correlated with latent growth at midyear ($r = -0.49$) and also spring ($r = -0.30$). This shows that examinees with lower initial latent ability grow more than examinees with higher initial latent ability.

Table 4.3

Fixed Effects of ABC Models

Fixed effect	ABC-M 1	S.E.	ABC-M7	S.E.	Odds ratio	ABC-M8	S.E.	Odds ratio
Intercept	--	--	3.97‡	0.25		3.28‡	0.39	
Time 2	2.33‡	0.04	2.49‡	0.09		2.65‡	0.09	
Time 3	4.08‡	0.07	4.56‡	0.15		4.64‡	0.16	
Shape 1 (NOFC)	4.16‡	0.24	--	--		--	--	
Shape 2 (SC)	2.67‡	0.20	-1.48‡	0.30		-0.88*	0.46	
Shape 3 (OFC)	2.32‡	0.21	-1.76‡	0.32		-1.14*	0.48	
Shape 4 (VOFC)	1.03‡	0.17	-3.06‡	0.29		-2.43‡	0.45	
Time2:Shape2			0.03	0.08		0.05	0.08	
Time3:Shape2			-0.29*	0.14		-0.22	0.15	
Time2:Shape3			-0.17*	0.08		-0.15	0.09	
Time3:Shape3			-0.61‡	0.14		-0.54†	0.15	
Time2:Shape4			-0.16*	0.08		-0.14	0.08	
Time3:Shape4			-0.41†	0.14		-0.33*	0.15	
Prek			0.49‡	0.07	1.63	0.73‡	0.08	
Disability			-1.15‡	0.14	0.32	-0.81‡	0.16	
ELL			-1.09‡	0.14	0.34	-1.49‡	0.16	
Age in fall			0.06‡	0.01	1.06	0.08‡	0.01	
Time 2: Prek						-0.38‡	0.06	0.68
Time 3: Prek						-0.48‡	0.09	0.62
Time 2: Disability						-0.45‡	0.11	0.64
Time 3: Disability						-0.56†	0.16	0.57
Time 2: ELL						0.51‡	0.12	1.67
Time 3: ELL						1.00‡	0.18	2.72
Time 2: Age in fall						-0.04‡	0.01	0.96
Time 3: Age in fall						-0.05‡	0.01	0.95

Fixed effect estimates of item properties in ABC-M1 represent the average easiness (on the logit scale) of items with each property. Interpretation of item easiness is that larger coefficients indicate items or item properties that are easier than those with

smaller coefficients. As shown by ABC-M1 estimates in Table 4.3, the very-often-confused property (VOFC) is most difficult with easiness of 1.03, followed by the often-confused property (OFC) with easiness of 2.32 and the sometimes-confused property (SC) with easiness of 2.67. The not-often-confused property (NOFC) with easiness of 4.16 was the easiest property for examinees. This ordered pattern of easiness indicates that a child's probability of recognizing the letter decreases as shape confusability increases. It also confirms expectations about letter shape confusability. Item properties of the ABC subtest did not account for all of the variance in item difficulty as indicated by a variance of 0.38 (see Table 4.4).

Table 4.4

Random Effects of ABC Models

				Person Correlations	
Model	Effect	Variance	SD	Time 1	Time 2
ABC-M1	Person Time 1	7.07	2.66		
	Person Time 2	2.14	1.46	-0.49	
	Person Time 3	1.65	1.28	-0.30	0.17
	Item Properties	0.38	0.62		
ABC-M7	Person Time 1	6.59	2.57		
	Person Time 2	2.14	1.46	-0.45	
	Person Time 3	1.71	1.31	-0.29	0.15
	Item Properties	0.39	0.62		
ABC-M8	Person Time 1	6.61	2.57		
	Person Time 2	2.03	1.42	-0.46	
	Person Time 3	1.63	1.28	-0.30	0.14
	Item Properties	0.45	0.67		

Models ABC-M2 to ABC-M5 were used to test the lack of time-invariance of each item property separately. In these four LG-EIRMs, the interaction effect of the particular item property by each time point was tested (see Table 4.5; De Boeck et al., 2011). Each of the four models was compared to ABC-M1 to evaluate the degree of lack of time-invariance (i.e. drift).

Table 4.5

Fixed Effects of Item Property Drift of ABC

Model	Fixed effect	Estimate	S. E.
ABC-M2	drift of NOFT at Time 2	0.09	0.08
	drift of NOFT at Time 3	0.40 ‡	0.14
ABC-M3	drift of SC at Time 2	0.18 ‡	0.04
	drift of SC at Time 3	0.14 *	0.06
ABC-M4	drift of OFC at Time 2	-0.08 *	0.04
	drift of OFC at Time 3	-0.26 ‡	0.06
ABC-M5	drift of VOFC at Time 2	-0.10 ‡	0.03
	drift of VOFC at Time 3	0.03	0.05

* p. 0.05
† p. < 0.01
‡ p. < 0.001

Table 4.6

Model Comparisons of ABC-M1 to ABC-M5

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	p value
ABC-M1	152501	152642	-76237	152475			
ABC-M2	152496	152659	-76233	152466	8.69	2	0.01
ABC-M3	152481	152644	-76226	152451	23.50	2	<0.001
ABC-M4	152483	152646	-76226	152453	21.67	2	<0.001
ABC-M5	152495	152658	-76233	152465	9.45	2	0.01

Table 4.5 shows that most interactions were statistically significant, which indicates that all four shape-confusability item properties showed lack of time-invariance. These results are confirmed by most model fit statistics (see Table 4.6). AIC, deviance, and LR test supported the more complex model, but BIC favored ABC-M1. More specifically, NOFC property only exhibited drift in spring and VOFC property only showed drift at mid-year. However, SC and OFC properties exhibited drift in both midyear and spring test administrations. It is interesting to look at the directions of the drift effects. NOFC and SC became easier because their interaction estimates were positive values. On the contrary, OFC and VOFC shifted to be more difficult since the values of their interaction estimates were negative. Thus, it seems that easy items become easier and difficult items became more difficult over time.

ABC-M6 also tested the drift effects (time by letter shape confusability interaction) but instead of testing each interaction separately, all interactions were included simultaneously. The purpose was to compare the fit of a model with all drift terms to a model without any drift terms (i.e. ABC-M1). All model fit criteria in Table 4.7 except BIC suggested ABC-M6 had a better fit than ABC-M1. Therefore, ABC-M6 was selected as the basis for subsequent model comparisons.

Table 4.7

Model Comparison of ABC-M1 and ABC-M6

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	P value
ABC-M1	152501	152642	-76237	152475			
ABC-M6	152464	152671	-76213	152426	48.072	6	<0.001

ABC-M7 was built upon ABC-M6 by including several person properties. The purpose of ABC-M7 was to investigate the influence of age and group membership on latent ability. Group effects included pre-k schooling, disability status, and ELL status. ABC-M7 estimates in Table 4.3 show that all four person properties were statistically significant. Students who attended pre-k school had scores that were an average of 0.49 logits (1.63 odds ratio) higher than the children who did not attend pre-k school. Children with disabilities were an average of 1.15 logits (0.32 odds ratio) lower than students without any disability. Similarly, the average latent ability of ELLs was 1.09 logits (0.34 odds ratio) lower than non-ELLs. Age in fall variable was centered to the grand mean of child age in fall 2013, so the value of 0.06 logits (1.06 odds ratio) indicates that performance increased as age increased. Random effect estimates shown in Table 4.4 suggest that person random effects and correlations between person estimates of ABC-M7 showed a very similar pattern to ABC-M1 even after controlling for the item property drift and overall person group difference in the model. That is, variance decreased over time and lower scoring students showed more growth over time.

ABC-M7 was compared to ABC-M6 to determine which model fits the data better. The result shown in Table 4.10 revealed that ABC-M7 was a better-fitting model based on all criteria. Thus, ABC-M7 was selected as the new basis for the next model comparison.

Table 4.8

Model Comparison of ABC-M6 and ABC-M7

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	P value
ABC-M6	152464	152671	-76213	152426			
ABC-M7	152265	152515	-76110	152219	207.29	4	<0.001

ABC-M8 is the most complex as it includes everything from ABC-M7 plus additional terms for the person property-by-time interactions. All interactions between time point and person properties were statistically significant (see Table 4.3). As such, we focus on reporting and interpreting the time-by-person property interactions in this model and ignore the main effects. As for pre-k schooling, the difference in the average amount of latent growth on the logit scale between children who attended pre-k school and children who did not was -0.38 logits (0.68 odds ratio) at midyear and -0.48 logits (0.62 odds ratio) in spring respectively. The direction of both interactions indicated that children who attended pre-k actually had less growth in alphabet recognition than children who did not attend pre-k. As for disability status, the latent growth of children with disability was 0.45 logits (0.64 odds ratio) smaller at mid-year and 0.56 logits (0.57 odds ratio) smaller in spring than it was for children without a disability. Looking at the interaction of time point by ELL status, ELLs had more growth than non-ELLs with the magnitude of 0.51 logits (1.67 odds ratio) at midyear and 1.00 logits (2.72 odds ratio) in spring. Age in fall is the only continuous person property variable in the model and the two interaction estimates for it were -0.04 logits (0.96 odds ratio) at midyear and -0.05 logits (0.95 odds ratio) in spring. This indicates that on average, older children grew less than younger children across time. With respect to person effects and correlations

between person estimates, random effect estimates of ABC-M8 shown in Table 4.4 indicate a very similar pattern to ABC-M1 and ABC-M7, even though item property drift and overall difference in latent ability and growth between person groups were accounted for in ABC-M8; children's scores showed less variability over time and low initial scores were related to more growth than high initial scores.

In Table 4.4, the variance component of items after controlling for the four item properties and item property-by-time point interactions was 0.45, which indicates that notable variance remained among ABC item difficulty even after the letter shape confusability properties and item property drift were taken into account.

Model comparison shows ABC-M8 fits the data significantly better than ABC-M7, according to all model fit criteria listed in Table 4.9, including LR test, deviance, AIC, and BIC. In other words, ABC-M8 is the best-fitting LG-EIRM for the ABC subtest.

Table 4.9

Model Comparison of ABC-M7 and ABC-M8

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	P value
ABC-M7	152265	152515	-76110	152219			
ABC-M8	152141	152479	-76040	152079	139.7	8	<0.001

Letter Sounds (LS). Model LS-M1 was the most basic LG-EIRM to be tested (see Table 4.10 and 4.11). With latent ability in fall as the baseline, the average latent growth in letter-sound knowledge was 2.88 at mid-year and 5.33 in spring on the logit scale, both of which were statistically significant. This suggests that, on average, a kindergartener grew significantly in LS from fall to midyear and continued to grow to

spring. Random effect estimates shown in Table 4.11 indicate a large amount of variance in fall ($\hat{\sigma}^2 = 7.11$), a smaller amount of variance at midyear ($\hat{\sigma}^2 = 2.23$), and even less variance at the spring time point ($\hat{\sigma}^2 = 1.74$), which suggests individual differences in LS decreased over time. The fixed and random effects also provide evidences of the appearance of a ceiling effect such that by the end of the year most students can pronounce all of the letter sounds. Looking at the correlations between variance components of persons, the latent ability in fall was negatively associated with the latent growth at midyear and in spring, with correlations of -0.49 and -0.16, respectively. This reveals that examinees with lower initial latent ability grow more than examinees with higher initial latent ability.

Table 4.10

Fixed Effects for LS Models

Fixed effect	LS-M1	S.E.	LS-M8	S.E.	Odds ratio	LS-M9	S.E.	Odds ratio
Intercept	--	--	0.23	0.41		-0.08	0.61	
Time 2	2.88‡	0.03	2.79‡	0.04		3.01‡	0.05	
Time 3	5.33‡	0.05	5.34‡	0.05		5.66‡	0.06	
Sound 1 (CV)	0.27	0.40	--	--		--	--	
Sound 2 (VC)	1.50†	0.43	1.18	0.77		1.01	0.10	
Sound 3 (VO)	-0.16	0.42	-0.34	0.69		-0.17	0.10	
Sound 4 (NA)	-0.16	0.43	-0.42	0.66		-0.23	0.10	
Sound 5 (Digraph)	-2.05‡	0.45	-2.46†	0.72		-1.58	1.16	
Time2:Sound2			0.20‡	0.05		0.20‡	0.04	
Time3:Sound2			0.08	0.07		0.08	0.07	
Time2:Sound3			-0.10*	0.04		-0.10†	0.04	
Time3:Sound3			-0.38‡	0.05		-0.37‡	0.05	
Time2:Sound4			0.12†	0.04		0.12†	0.04	
Time3:Sound4			-0.36‡	0.05		-0.35‡	0.05	
Time2:Sound5			0.08	0.05		0.09	0.05	
Time3:Sound5			0.40‡	0.06		0.41‡	0.06	
Prek			0.35‡	0.07	1.42	0.70‡	0.08	
Disability			-1.37‡	0.14	0.25	-0.10‡	0.16	
ELL			-1.08‡	0.14	0.34	-1.56‡	0.16	
Age in fall			0.06‡	0.01	1.06	0.09‡	0.01	
Time 2: Prek						-0.44‡	0.05	0.64
Time 3: Prek						-0.66‡	0.07	0.52
Time 2: Disability						-0.43‡	0.10	0.65
Time 3: Disability						-0.71†	0.14	0.49
Time 2: ELL						0.59‡	0.11	1.80
Time 3: ELL						0.87‡	0.15	2.39
Time 2: Age in fall						-0.04‡	0.01	0.97
Time 3: Age in fall						-0.05‡	0.01	0.95

* p. 0.05
† p. < 0.01
‡ p. < 0.001

As for item properties, fixed effect estimates in Table 4.10 represent the average easiness of the five LS item properties on the logit scale. Comparing them to each other, the Digraph property is most difficult with easiness of -2.05, followed by NA and VO with easiness of -0.16 for both. The CV and VC were easier than the other three properties with easiness of 0.27 and 1.50 respectively. Among them, VC turned out to be the easiest property for examinees. This suggests that a child has the highest probability of correctly recognizing the sounds of letters with VC name structure and the lowest probability of correctly identifying the sounds of letters with Digraph name structure. The variance of items shown in Table 4.11 was 2.66, which suggests there was a large amount of variance among LS items even after accounting for the five item properties.

Table 4.11

Random Effects for LS Models

				Person Correlations	
Model	Effect	Variance	SD	Time 1	Time 2
LS-M1	Person Time 1	7.11	2.67		
	Person Time 2	2.23	1.49	-0.49	
	Person Time 3	1.74	1.32	-0.16	0.14
	Item Properties	2.66	1.63		
LS-M8	Person Time 1	6.76	2.60		
	Person Time 2	2.59	1.50	-0.45	
	Person Time 3	1.66	1.29	-0.22	0.13
	Item Properties	2.69	1.64		
LS-M9	Person Time 1	6.74	2.60		
	Person Time 2	2.16	1.47	-0.46	
	Person Time 3	1.62	1.27	-0.22	0.10
	Item Properties	2.95	1.72		

Models LS-M2 to LS-M6 were used to test a lack of time-invariance of each item property of LS items separately. In each of these five LG-EIRMs, the interaction effects of a particular item property by time points were created and tested (De Boeck et al., 2011). Considering the main purpose of these models, only estimates of those interaction effects were of interest and reported in Table 4.12. Additionally, each of the five models was compared to LS-M1 to evaluate the degree of the lack of time-invariance of item properties (i.e. drift). The model comparison results were listed in Table 4.13.

Table 4.12

Fixed Effects of Item Property Drift of LS

Model	Fixed effect	Estimate	S. E.
LS-M2	drift of CV at Time 2	-0.04	0.03
	drift of CV at Time 3	0.08 *	0.04
LS-M3	drift of VC at Time 2	0.19 ‡	0.04
	drift of VC at Time 3	0.22 ‡	0.06
LS-M4	drift of VO at Time 2	-0.17 ‡	0.04
	drift of VO at Time 3	-0.38 ‡	0.06
LS-M5	drift of NA at Time 2	0.11 ‡	0.03
	drift of NA at Time 3	-0.36 ‡	0.04
LS-M6	drift of Digraph at Time 2	0.06	0.04
	drift of Digraph at Time 3	0.58 ‡	0.05

* p. 0.05
† p. < 0.01
‡ p. < 0.001

From Table 4.12, we can see that most item property–by-time point interactions were statistically significant and all five letter-name structure properties showed different degrees of lack of time-invariance. More specifically, CV and Digraph properties only

exhibited drift in spring test and the other three properties exhibited drift in both midyear and spring test administrations. Looking at the directions of the drift effects, CV, VC, and Digraph became easier because their interaction estimates were positive values. In contrast, VO shifted to be more difficult since the values of its interaction estimates were below 0 in both midyear and spring tests. As for NA, the result is mixed as the directions of its drift were not consistent across time. Besides, Table 4.13 showed that multiple model fit criteria (i.e. AIC, LR test, and deviance) indicated LS-M2 to LS-M6 fit better to the data than LS-M1, although BIC favored LS-M1 over LS-M2.

Table 4.13

Model Comparisons of LS-M1 to LS-M6

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	p value
LS-M1	210207	210359	-105090	210179			
LS-M2	210201	210375	-105085	210169	9.52	2	0.01
LS-M3	210182	210356	-105075	210150	28.83	2	<0.001
LS-M4	210131	210304	-105049	210099	80.50	2	<0.001
LS-M5	210099	210273	-105034	210067	111.80	2	<0.001
LS-M6	210044	210218	-105006	210012	167.01	2	<0.001

Model LS-M7 included all interactions of time point by item property simultaneously. It aimed to compare the fit of a model with all drift terms to a model without any drift terms (i.e. LS-M1). Table 4.14 shows LS-M7 had better fit than LS-M1. Therefore, LS-M7 was selected as the basis for subsequent model comparisons.

Table 4.14

Model Comparison of LS-M1 and LS-M7

Model	AIC	BIC	LogLike lihood	Deviance	Chi sq	df	P value
LS-M1	210207	210359	-105090	210179			
LS-M7	209902	210142	-104929	209858	320.53	8	<0.001

In model LS-M8, fixed effects of person properties were all statistically significant (see Table 4.10), indicating all four person properties were significantly related to the individual differences in latent ability of letter sounds among kindergarteners. As for pre-k schooling, the average latent ability of children who attended pre-k school was 0.35 logits (1.42 odds ratio) higher than the children who did not attend pre-k school. Children who have a disability were 1.37 logits (0.25 odds ratio) lower on average than those without a disability. Similarly, the average latent ability of ELLs was 1.08 logits (0.34 odds ratio) lower than non-ELLs. As a continuous variable that was centered to the grand mean, the estimate with value of 0.06 logits (1.06 odds ratio) of age in fall indicates on average, older children have higher latent ability than younger children. Random effect estimates of LS-M8 shown in Table 4.11 suggest that its person random effects and correlations between person estimates of showed a very similar pattern to LS-M1 even after controlling for the item property drift and overall person group difference in the model.

Table 4.15 shows that LS-M8 fit the data better than LS-M7, based on all criteria that include AIC, BIC, LR test, and deviance. Thus, LS-M8 was selected as the new basis for the next model comparison.

Table 4.15

Model Comparison of LS-M7 and LS-M8

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	P value
LS-M7	209902	210142	-104929	209858			
LS-M8	209686	209969	-104817	209634	224.58	4	<0.001

Model LS-M9 is the most complex and general LG-EIRM to be tested for LS. Table 4.10 presents fixed effect estimates of LS-M9 which show all interactions between time point and person properties were statistically significant. Thus, we only focus on reporting and interpreting those time-by-person property interactions in this model using average latent ability in fall as the reference level. As for pre-k schooling, the difference in average amount of latent growth between children who attended pre-k school and children who did not was -0.44 logits (0.64 odds ratio) at midyear and -0.66 logits (0.52 odds ratio) in spring, respectively. The directions of these interactions indicated that children who attended pre-k school actually grew less in letter sounds than children who did not attend pre-k school. As for disability, compared to children without a disability, the average latent growth of children with a disability was 0.43 logits (0.65 odds ratio) smaller at midyear and 0.71 logits (0.49 odds ratio) smaller in spring. Looking at the interaction of time point by ELL status, we find ELLs grew more in letter sounds than non ELLs, with a magnitude of 0.59 logits (1.80 odds ratio) at midyear and 0.87 logits (2.39 odds ratio) in spring. Age in fall is the only continuous person property variable in the model and the two interaction estimates of it were -0.04 logits (0.97 odds ratio) at midyear and -0.05 logits (0.95 odds ratio) in spring. This indicates older children grew less than younger children across time points. In regard to its person effects and their

correlations, random effect estimates of LS-M9 shown in Table 4.11 indicate a very similar pattern to LS-M1 and LS-M8, even though item property drift and overall difference in latent ability and growth between person groups were controlled in the model; variance in latent growth decreased over time and lower initial scores were related to more growth at fall and spring. Item properties of LS did not account for all of the variance in item difficulty as indicated by a variance of 2.95 (see Table 4.11).

Model comparison shows that LS-M9 fit the data significantly better than LS-M8, according to all criteria listed in Table 4.16, including LR test, deviance, AIC, and BIC. In other words, LS-M9 is the best-fitting LG-EIRM for LS. This confirms the results from analyzing the fixed effects of LS-M9 that all four person properties have significant impact on the average latent growth of examinees in LS.

Table 4.16

Model Comparison of LS-M8 and LS-M9

Model	AIC	BIC	Log Likelihood	Deviance	Chi sq	df	P value
LS-M8	209686	209969	-104817	209634			
LS-M9	209494	209864	-104713	209426	208.06	8	<0.001

Chapter 5 Discussion

The present study applied an explanatory item response model approach to examine kindergartener's latent growth in two aspects of alphabet knowledge: alphabet recognition and letter sounds. Item- and person-relevant factors associated with child's latent ability and latent growth in alphabet knowledge were also investigated. This chapter answers the research questions and discusses the implications of results.

Growth

The results of the study revealed that kindergarteners grew significantly in both alphabet recognition and letter sound knowledge. The largest amount of growth occurred from fall to midyear, but growth continued to spring.

Figure 5.1 illustrates the variability in kindergartener's growth across three time points. Not all children showed the same magnitude or pattern of growth. As shown in the figure, student scores improve over time but for some the improvement is linear while it is quadratic for others. Moreover, variability among scores decreases over time. This result is also illustrated in Figure 5.1 and the random effects from the various LG-EIRMs.

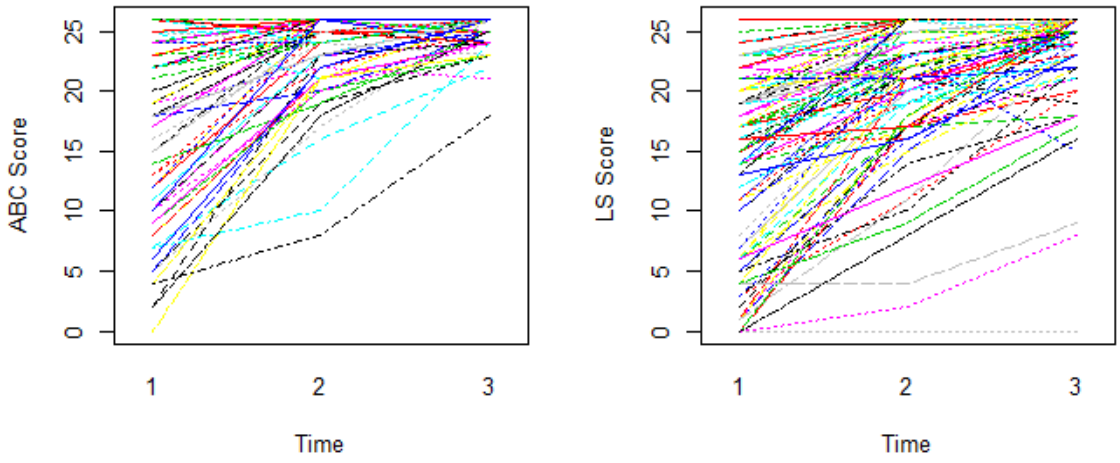

Figure 5.1 Sum score-based growth trajectories of a 100 random sample

Factors Influencing Growth

By incorporating several person properties into the LG-EIRMs, the study was able to compare different person groups in terms of their overall ability and growth. The results from LG-EIRM growth analysis suggested all four person properties examined in this study, including pre-k schooling, disability status, ELL status, and child's age in fall, were statistically significant in models for both ABC and LS subtests. To give a more intuitive sense of the results, Figure 5.2 shows average growth trajectories for both ABC and LS subtests.

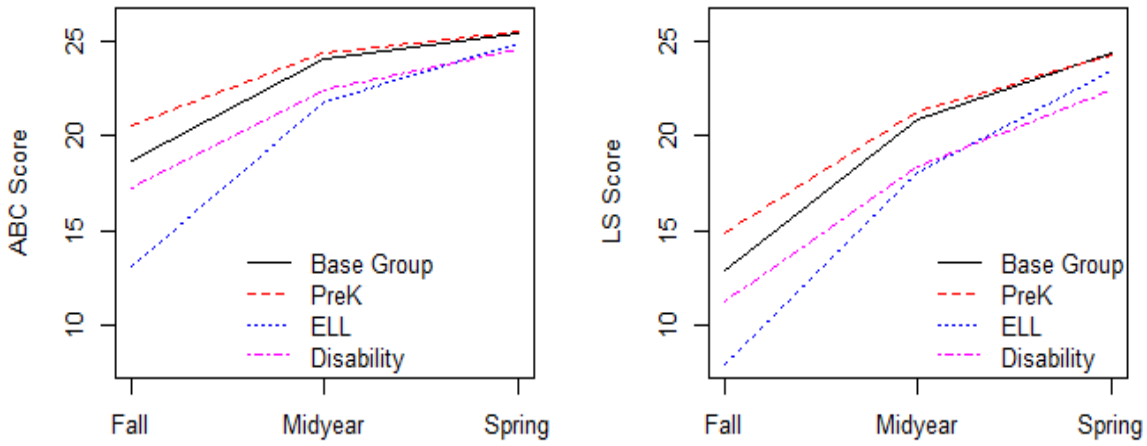

Figure 5.2 Average growth trajectories of multiple person groups

As for pre-k schooling, examinees who attended pre-k school had significantly higher overall latent ability level in alphabet knowledge than examinees who did not, as suggested by estimates of ABC-M7 and LS-M8. However, when it comes to latent growth, examinees with pre-k schooling grew significantly less over the year than examinees without pre-k schooling as shown by estimates of ABC-M8 and LS-M9. In Figure 5.2, the slope of the non-pre-k group was steeper than the pre-k group, indicating the non-pre-k group grew more. However, the pre-k group scored higher than the non-pre-k group, especially at the first time point. Regarding the differences, there are few explanations. Most children who attend pre-k school are taught letter names and letter sounds which are constrained skills (Paris, 2005). Thus, they have already mastered a large portion of the 26 alphabet letters before attending kindergarten leading to better initial performance in fall than children without pre-k school experience (Bear, Invernizzi, Templeton, & Johnston, 2000). Due to the ceiling effect, they had less room to grow than children who never attended pre-k school and thus their performance scores

appeared to increase less. From Figure 5.2, we also see that it approximately took until midyear for non-pre-k children to catch up with pre-k children in terms of their overall performance. Thus, children with pre-k school experience are prepared for early literacy skills sooner because they are about a half of a school year ahead of children without pre-k school experience.

According to LG-EIRM analysis, examinees with a disability had significantly lower overall ability and a significantly smaller amount of latent growth in alphabet knowledge than examinees without a disability. Earlier studies have shown that differences between children with a reading disability and normal children in alphabet knowledge measure diminish over kindergarten (Smith et al., 2008; Compton et al., 2006). Our results also show that the difference gets smaller, but non-disabled children show more growth and a gap between disabled and non-disabled children persists until the end of kindergarten. What suggested by the plots in Figure 5.2 confirmed the information we obtained from LG-EIRM analysis. In Figure 5.2, the average growth trajectory of the disability group is clearly located below the trajectory of no-disability group and the slope of the average growth trajectory of the disability group was detectably smaller than that of disability group. The reason behind the fact that disabled children scored lower initially and grew less than non-disabled children is clear. In our random sample, disabled children with disability mostly fell into the three largest categories: learning disability, speech/language impairment, and developmental delay. So it was expected that children with those types of disabilities experienced difficulties in developing their alphabet knowledge and achieving high scores.

Growth analysis results for ELL students were similar to those for students without pre-k schooling. ELLs had significantly lower overall latent ability level in alphabet knowledge than non-ELLs, but they grew significantly more over time than non-ELLs (see Figure 5.2). However, non-ELLs still outperformed ELLs in terms of overall ability in alphabet recognition and letter sounds. As for the difference between ELLs and non-ELLs, there could be two primary sources. Children who are non-ELLs have better knowledge of and more exposure to alphabet knowledge at home because their parents teach them letters via a variety of daily activities and talk to them in English. On the contrary, it is very likely that ELLs speak their first languages other than English at home. So ELLs scored lower initially in fall but had more room to grow, compared to non-ELLs. Another important reason is the bilingual advantage in metalinguistic development (Vygotsky, 1962; Campbell & Sais, 1995). A sizable amount of studies found exposure to the second language may help improve children's development of metalinguistic skills (Hakuta, 1986; Campbell & Sais, 1995; Galambos & Hakuta, 1988). Particularly, being bilingual benefits children's speech-sound awareness development (Campbell & Sais, 1995). The results from this study showed evidence that being bilingual may bring benefits to kindergarteners during their developmental progress of alphabet knowledge.

As for the only continuous person property child's age in fall, the estimates from LG-EIRM analysis revealed that older children grew significantly less than younger children. Therefore, a child's age in fall has a significant impact on the amount of growth of his/hers in alphabet knowledge. However, older children had better performances in both alphabet recognition and letter sounds than younger children, which corroborates

findings from some previous studies that suggested upon kindergarten entry, older child outperform younger children on average (Oshima & Domaleski, 2006; Stipek, 2002; Huang & Invernizzi, 2012). The way that age influences child's growth in alphabet knowledge is similar to pre-k schooling. Older children got longer exposure to letters and know more letter names and letter sounds before attending kindergarten, so they performed significantly better initially but appeared to grow less over time than younger children. Accordingly, the early-age achievement gap will narrow down over time. However, it will not completely disappear (Huang & Invernizzi, 2012).

Item properties and their lack of time-invariance

Results showed that the degree of letter-shape confusability was ordered. NOFC property appeared to be the easiest, followed by SC and OFTC in turn. VOFC property was the most difficult. This order is expected because the more distinguishable the shape, the easier it is for child to learn and recognize it.

LS item properties did not show the same strict ordering as did ABC item properties. VC and CV properties were the two easiest properties, followed by VO and NA properties. Digraph was determined to be the most difficult. In general, this is consistent with the findings from prior studies that letter sounds which are more tightly associated to their names are more easily for early-age children to learn and master (McBride-Chang, 1999; Evans et al., 2006; Huang, Tortorelli, & Invernizzi, 2014). However, specifically regarding the ordering of VC and CV, this study showed that CV letter sounds are more difficult than VC letter sounds, which is not consistent with previous studies (Huang, Tortorelli, & Invernizzi, 2014; Evans et al., 2006). However,

Huang and Invernizzi's (2014) controlled for the letter frequency in the English language in their study when comparing the difficulty of CV and VC letters, whereas this study did not . Controlling for letter frequency may have led to a different ordering of CV and VC difficulties. More specifically, the median frequency for CV letters is 18.55 and the median frequency for VC letters is 26.10. This indicates VC letters are more common than CV letters, which may make VC letters easier to learn because of more exposure to them of children.

Furthermore, the results from models testing item property drift showed that most ABC item properties and LS item properties exhibited different degrees and directions of lack of time-invariance. The causes would certainly not be from children's ability growth because the growth has been accounted for in all LG-EIRMs. A few possible explanations could be children's perceptions of items, practice effect of children at school and home, and teachers' emphasis on particular letters that they believe to be more difficult during instruction. For instance, from ABC models, we found easy item properties (i.e. NOFT, SC) drifted to be even easier and hard item properties (i.e. OFC, VOFT) became more difficult over time. This might be due to children having more confidence and more self-efficacy for items they recognize and perceive as easy. Conversely, children may show lack of confidence and low self-efficacy for letters they recognize and perceive as difficult, thereby making difficult letters more difficult to recognize. Another example is that LS models showed the Digraph property drifted to become significant easier in spring. The reason might be that teachers believed the sounds of digraphs were generally harder than that of single letters and such that put stronger emphasis and allocate more time on digraphs during their instructions. Thus, children had more practice with digraph

items and as a result, most children learned to pronounce them correctly at the end of school year. In addition, VC property was found to become easier across time which could be due to that children being more confident and showing high self-efficacy for VC letters which they recognized as the easiest.

Implications

Alphabet knowledge is deemed to be one of the core early literacy skills, which mainly includes two components, letter-name knowledge and letter-sound knowledge (Invernizzi, 2004; Huang, Tortorelli, & Invernizzi, 2014). Previous studies indicated alphabet knowledge is a strong predictor of child's future reading performance and contributes greatly to child's success in future reading and writing (Adams, 1990; Foulin, 2005; Snow, et al., 1998; Stevenson & Newman, 1986). Although numerous studies have examined the role of alphabet knowledge as a predictor of a child's later reading performance (Dickinson, Tabors, & Roach, 1996; Scarborough 1998; Lonigan, Burgess, and Anthony, 2000; Missall & McConnell, 2010; NELP, 2008), studies rarely investigate the development of alphabet knowledge over time. To promote understandings of a child's growth in alphabet knowledge, the present study focused on measuring kindergarteners' growth in both alphabet recognition and letter sounds and explored various factors that influence such growth. The results of the study revealed that kindergarteners grow significantly in alphabet recognition and letter sounds, from fall to spring, but there is a lot of variation among overall performance and growth trajectory. This confirms that kindergarten is a crucial period for child to develop knowledge of alphabet letters (VanDerHeyden, Snyder, Broussard, & Ramsdell, 2008) and also

suggests that it is essentially important to examine concurrent development of children in early literacy skills if multiple-time-point assessment data are available for one school year.

Furthermore, the present study indicates access to item-level early literacy assessment data has significant practical meanings to tracking a child's learning progress and exploring the influences of different factors on literacy growth of children. Unlike the majority of previous growth studies that used task scores or sum scores to estimate growth (Pan, et al., 2005; McCoach,et al. 2006; Hammer, Lawrence, & Miccio, 2007; Connor, Morrison, & Slominski, 2006; Gutierrez & Vanderwood, 2013), the present study used item-level responses and an IRT approach to analyze growth on the scale of the latent construct directly. The impact from factors that are related to items and persons on the latent growth was also analyzed and the information of such impact can be used for practical purposes, such as pinpointing student understanding in various meaningful ways. For example, the item property estimates can be used to connect child's initial score and change scores to the certain group of items and thus associate child's probability of correctly answering the letter to the specific feature that the letter has (Cho et al., 2013). In addition, the information about the person properties as the sources of individual differences in growth can be taken advantage of to guide customization of instructional support.

Moreover, the sample size used in this study is larger than most previous longitudinal studies of early literacy (Connor, Morrison, & Slominski, 2006; Ding, 2012; Morris, Bloodgood, Lomax, & Perney, 2003; Wagner, et al., 1997), which strengthened

the validity of its results and inferences. Another notable feature of it is the inclusion and investigation of various person properties (i.e. pre-k schooling, disability, ELL status, age) in one study which has rarely be done by other growth studies.

However, this study has few limitations. Due to the model complexity, this study only used a 5,000 random sample instead of the full population of examinees. Additionally, the absolute model fit of LG-EIRMs have not been evaluated in this study. These limitations can be used to provide possible directions of future studies.

References

Adams, M. (1990). *Beginning to read: Thinking and learning about print*. Urbana Champaign, IL: University of Illinois.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1-23.

Adams, R. J., Wilson, M., & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47–76.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (pp. 199-213). New York: Springer

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3-16.

Arciuli, J., & Simpson, I. C. (2011). Not all letters are created equal: Exploring letter name knowledge through spelling in school children and adults. *Writing Systems Research,* 3, 1–13.

Bacci, S. (2012). Longitudinal data: different approaches in the context of item-response theory models. *Journal of Applied Statistics*, *39*(9), 2047-2065.

Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F (2000). *Words their way: Word study for phonics, vocabulary, and spelling instruction* ((2 nd ed.)). Upper Saddle River, NJ: Merrill.

Betebenner, D.W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*, 42-51.

Betebenner, D., & Linn, R.. (2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Retrieved June, 25, 2014.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Bock, R., Muraki, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models* (Tech. Rep.). Princeton, New Jersey: Educational Testing Service.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89-118.

Briggs, D., & Betebenner, D. (2009). *Is growth in student achievement scale dependent?* Paper presented at the annual meeting of National Council for Measurement in Education, San Diego, CA.

Briggs, R., & Hocevar, D. J. (1975). A new distinctive feature theory for upper case letters. *Journal of General Psychology*.

Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, *62*(1), 23-43.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*(1), 171-189.

Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: An introduction to the latent growth curve model. *Understanding Statistics*, *2*(3), 177-203.

Cabell, S. Q., Justice, L. M., Konold, T. R., & McGinty, A. S. (2011). Profiles of emergent literacy skills among preschool children who are at risk for academic difficulties. *Early Childhood Research Quarterly, 26*, 1–14.

Camilli, G. (2006). Test fairness. *Educational measurement, 4,* 221-256.

Campbell, R., & Sais, E. (1995). Accelerated metalinguistic (phonological) awareness in bilingual children. *British Journal of Developmental Psychology*,*13*(1), 61-68.

Castellano, K. E., & Ho, A. D. (2013a). *A practitioner's guide to growth models.* Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*, 190-215.

Chaney, C. (1992). Language development, metalinguistic skills, and print awareness in 3-year-old children. *Applied Psycholinguistics, 13*, 485-514.

Cho, S. J., Gilbert, J. K., & Goodwin, A. P. (2013). Explanatory Multidimensional Multilevel Random Item Response Model: An Application to Simultaneous Investigation of Word and Person Contributions to Multidimensional Lexical Representations. *Psychometrika, 78*(4), 830-855.

Cho, S. J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 353-381.

Cohn, M., & Stricker, G. (1976). Inadequate perception vs. reversals. *The Reading Teacher*, 162-167.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. Journal of Education Psychology, 98, 394–409.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305-336.

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology*, *98*(4), 665.

Coyne, M. D., & Harn, B. A. (2006). Promoting beginning reading success through meaningful assessment of early literacy skills. *Psychology in the Schools*, *43*(1), 33-43.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*(2), 147-167.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1-28.

Deno, S. L., & Fuchs, L. S. (1987). Developing Curriculum-Based Measurement Systems for Data-Based Special Education Problem Solving. *Focus on Exceptional Children, 19*(8), 1-16.

Diamond, K. E., Gerde, H. K., & Powell, D. R. (2008). Development in early literacy skills during the pre-kindergarten year in Head Start: Relations between growth in children's writing and understanding of letters. *Early Childhood Research Quarterly, 23*(4), 467-478.

Dickinson, D.K., Tabors, P.O., & Roach, K.A. (1996). Contribution of early oral language skills to later reading comprehension. In A. Spinollo & J. Oakhill, *Thinking about texts: Comprehension and metalinguistic awareness*. Symposium conducted at the XIVth Biennial Meetings of ISSBD in August 1996. Quebec City, Quebec.

Dickinson, D., & Neuman, S. (Eds.). (2006). *Handbook of early literacy research* (Vol. 2). New York: Guilford.

Ding, C. (2012). Studying children's early literacy development: Confirmatory multidimensional scaling growth modeling. *International Journal of Educational Research, 53*, 278-288.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2), 1-18.

Duncan, T. E., & Duncan, S. C. (2004). An introduction to latent growth curve modeling. *Behavior therapy*, *35*(2), 333-363.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge Academic.

Ehri, L. C, & Roberts, T. (2006). The roots of learning to read and write: Acquisition of letters and phonemic awareness. *Handbook of early literacy research, 2,* 113-131.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495-516.

Embretson, S. E. (1995). Measurement Model for Linking Individual Learning to Processes and Knowledge: Application to Mathematical Reasoning. *Journal of educational measurement*, *32*(3), 277-294.

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

Evans, M., Bell, M., Shaw, D., Moretti, S., & Page, J. (2006). Letter names, letter sounds and phonological awareness: An examination of kindergarten children across letters and of letters across children. *Reading and Writing, 19,* 959–989.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, *58*(3), 357-381.

Fieuws, S., & Verbeke, G. (2004). Joint modeling of multivariate longitudinal profiles: pitfalls of the random effects approach. *Statistics in Medicine, 23,* 3093–3104.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course,* 269-314.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359-374.

Fiset, D., Blais, C., Ethier-Majcher, C., Arguin, M., Bub, D., & Gosselin, F. (2008). Features for identification of uppercase and lowercase letters.*Psychological Science*, *19*(11), 1161-1168

Flanigan, K. (2007). A concept of word in text: A pivotal event in early reading acquisition. *Journal of Literacy Research*, *39*, 37–70.

Fraine, B. D., Damme, J. V., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology*, *32*(1), 132-150.

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing: An Interdisciplinary Journal*, *18*, 129–155.

Galambos, S. J., & Hakuta, K. (1988). Subject-specific and task-specific characteristics of metalinguistic awareness in bilingual children. *Applied Psycholinguistics*, *9*(02), 141-162.

Good, R. H., & Kaminski, R. (2002). *Dynamic Indicators of Basic Early Literacy Skills 6th Edition (DIBELS)*. Eugene, OR: Institute for the Development of Educational Achievement. Available at http:///dibels.uoregon.edu.

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2011). DIBELS Next technical manual.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 369-377.

Grady, M., Lewis, D., & Gao, F. (2010). *The effect of sample size on student growth percentiles.* Monterey, CA: CTB/McGraw-Hill. Retrieved from http://www.ctb.com/ctb.com/control/researchArticleMainAction?articleId=17020.

Gutierrez, G., & Vanderwood, M. L. (2013). A growth curve analysis of literacy performance among second-grade, Spanish speaking, English language learners. *School Psychology Review, 42* (1), 3–21.

Hakuta, K. (1986). *Mirror of language: The debate on bilingualism*. Basic Books Inc., New York.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(2), 38-47.

Hammer, C. S., Lawrence, F. R., & Miccio, A. W. (2007). Bilingual children's language abilities and early reading outcomes in Head Start and kindergarten. *Language, Speech, and Hearing Services in Schools, 38*(3), 237-248.

Hammill, D. D. (2004). What we know about correlates of reading. *Exceptional Children, 70*, 453–468.

Han, K.T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. *GMAC Research Reports, RR-11-02.* Retrieved from: http://www.gmac.com/market-intelligence-and-research/research-library/validity-and-testing/research-reports-validity-related/potential-impact-of-item-parameter-drift-due-to-practice-and-curriculum-change-on-item-calibration.aspx

Heckman, J. J., & Masterov, D. V. (2007). The productivity argument for investing in young children. *Applied Economic Perspectives and Policy*, *29*(3), 446-493.

Helman, L. A. (2005). Using Literacy Assessment Results to Improve Teaching for English-Language Learners. *The Reading Teacher*, *58*(7), 668-677.

Hindman, A. H., Skibbe, L. E., Miller, A., & Zimmerman, M. (2010). Ecological contexts and early learning: Contributions of child, family, and classroom factors during Head Start, to literacy and mathematics growth through first grade. *Early Childhood Research Quarterly*, *25*(2), 235-250.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Routledge.

Huang, F., & Invernizzi, M. (2012). The case for confusability and other factors associated with lowercase alphabet recognition, applied psycholinguistics. *Advance Online Publication*, http://dx.doi.org/10.1017/S0142716412000604

Huang, F. L., & Konold, T. R. (2014). A latent variable investigation of the Phonological Awareness Literacy Screening-Kindergarten assessment: Construct identification and multigroup comparisons between Spanish-speaking English-language learners (ELLs) and non-ELL students. *Language Testing*, *31*(2), 205-221.

Huang, F. L., Tortorelli, L. S., & Invernizzi, M. A. (2014). An investigation of factors associated with letter-sound knowledge at kindergarten entry. *Early Childhood Research Quarterly*, *29*(2), 182-192.

Invernizzi, M., Juel, C., Swank, L., & Meier, J. (2011). Phonological Awareness Literacy Screening (PALS): K technical reference manual. *Charlottesville, VA: University of Virginia Curry School of Education*.

Invernizzi, M., Justice, L., Landrum, T., & Booker, K. (2004). Early literacy screening in

kindergarten: Widespread implementation in Virginia. *Journal of Literacy Research*, *36*, 479–500.

Invernizzi, M., Sullivan, A., Meier, J., & Swank, L. (2001). Phonological awareness literacy screening for preschool. *Virginia: The Rector and the Board of Visitors of the University of Virginia*.

Invernizzi, M., Sullivan, A., Meier, J., & Swank, L. (2004). Phonological awareness literacy screening-PreK. *Charlottesville, VA: University of Virginia*.

Janssen, R., & De Boeck, P. (2006). *A random-effects version of the LLTM*. Technical report, Department of Psychology, University of Leuven, Belgium.

Justice, L. M., Invernizzi, M. A., & Meier, J. D. (2002). Designing and implementing an early literacy screening protocol: Suggestions for the speech-language pathologist. *Language, Speech, and Hearing Services in Schools, 33*, 84-101.

Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, *99*(465), 279-290.

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

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (Vol. 10). Sage press.

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-scale Assessments in Education*, *2*(1), 1.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., et al. (1996). The reliability and validity of a new psychotherapy outcome questionnaire. *Clinical Psychology and Psychotherapy, 3*, 249-258.

Lambert, R. G. (2012). *Teaching Strategies GOLD® assessment system: Growth norms technical report*. Charlotte, NC: Center for Educational Measurement and Evaluation.

Levin, I., & Ehri, L. C. (2009). Young children's ability to read and spell their own and classmates' names: The role of letter knowledge. *Scientific Studies of Reading*, *13*(3), 249-273.

Liberman, I. Y., Shankweiler, D., Fischer, F., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*, 201–212.

Linn, R. L. (2008). Educational accountability systems. In *The future of test-based educational accountability* (pp. 3–24). New York: Taylor & Francis.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Developmental psychology*, *36*(5), 596.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

Lundberg, I., Frost, J., & Petersen, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading research quarterly*, 263-284

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quarterly (1982-)*, 285-308.

McBride-Chang, C. (1998). The development of invented spelling. *Early Education and Development*, *9*(2), 147-160.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology*, *98*(1), 14.

McCormick, C. E., & Haack, R. (2010). Early literacy individual growth and development indicators (EL-IGDIs) as predictors of reading skills in kindergarten through second grade. *Tarptautinis psichologijos žurnalas: Biopsichosocialinis požiūris*, (7), 29-40.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

McCulloch, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Pub. Co.

McGee, L. M., Lomax, R. G., & Head, M. H. (1988). Young children's written language knowledge: What environmental and functional print reading reveals.*Journal of Literacy Research*, *20*(2), 99-118.

McGuire, L. W. (2010). *Practical formulations of the latent growth item response model*. University of California, Berkeley.

McCulloch, C. E., & Neuhaus, J. M. (2001). *Generalized linear mixed models*. John Wiley & Sons, Ltd.

McCulloch, C. E., & Searle, S. R. (2001). Linear Mixed Models (LMMs). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*(2), 289-311.

Messick, S. (1990). Validity of test interpretation and use. Research report, retrieved from: http://files.eric.ed.gov/fulltext/ED395031.pdf

Meyer, J. P. (2010). *Reliability.* Oxford: Oxford University Press.

Meyer, J. P. (2014). *Applied Measurement with JMetrik*. Routledge.

Missall, K. N., & McConnell, S. R. (2010). Early literacy and language IGDIs for preschool-aged children. *Individual Growth and Development Indicators: Tools for monitoring progress and measuring growth in very young children*, 181-201.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1963). Introduction into the theory of statistics.McGraw-Hill, New York.

Morris, D., Bloodgood, J. W., Lomax, R. G., & Perney, J. (2003). Developmental steps in learning to read: A longitudinal study in kindergarten and first grade. *Reading research quarterly*, *38*(3), 302-328.

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first-and second-grade reading achievement. *The Elementary School Journal*, 93-109.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30.

Muter, V., & Diethelm, K. (2001). The contribution of phonological skills and letter knowledge to early reading development in a multilingual population. *Language Learning*, *51*, 187-219.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*(1), 43-65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

NAEYC. (1991). Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8. *Young Children, 30*, 21-38.

National Early Literacy Panel. (2008). *Developing early literacy: The report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Northwest Evaluation Association. (2011). *RIT scale norms study*. Portland, OR: Thum and Hauser.

Oshima, T. C., & Domaleski, C. S. (2006). Academic performance gap between summer-birthday and fall-birthday children in grades K-8. Journal of Educational Research, 4, 212–217.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families.*Child development*, *76*(4), 763-782.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading research quarterly*, *40*(2), 184-202.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied psychological measurement*, *30*(2), 100-120.

Pearson. (2012a). *aimsweb: Technical Manual*. Bloomington, MN.

Pearson. (2012b). *AIMSweb: ROI growth norms guide*. Bloomington, MN.

Piasta, S. B., & Wagner, R. K. (2010). Learning letter names and sounds: Effects of instruction, letter type, and phonological processing skill. *Journal of experimental child psychology*, *105*(4), 324-344.

Ponocny, I. (2002). On the applicability of some IRT models for repeated measurement designs: Conditions, consequences, and goodness-of-fit tests. *Methods of Psychological Research Online*, *7*(1), p22-40.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent Growth Curve Modeling*. Quantitative Applications in the Social Sciences.

Puranik, C. S., Lonigan, C. J., & Kim, Y. S. (2011). Contributions of emergent literacy skills to name writing, letter writing, and spelling in preschool children. *Early Childhood Research Quarterly, 26*, 465–474. http://dx.doi.org/10.1016/j.ecresq.2011.03.002

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests.

(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Raudenbush, S. W., & Bryk, A. G. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Renaissance Learning. (2012). *STAR Reading: Technical manual*. Wisconsin Rapids, WI.

Reutzel, D. R., Jones, C. D., & Clark, S. K. (2012). Teaching Text Structure to Improve Young Students' Knowledge Acquisition and Comprehension.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39*(5), 406-412.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*, 103-116.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research, 49*, 264-276.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether item parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64*, 588-599.

Rupp, A. A. & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66* (1), 63-83.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities:

Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–119). Timonium, MD: York Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Share, D. L. (2004). Knowing letter names and learning letter sounds: A causal connection. *Journal of Experimental Child Psychology, 88*, 213–233.

Smith, S. L., Scott, K. A., Roberts, J., & Locke, J. L. (2008). Disabled readers' performance on tasks of phonological processing, rapid naming, and letter knowledge before and after kindergarten. *Learning disabilities research & practice*, *23*(3), 113-124.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Speece, D. L., Ritchey, K. D., Cooper, D. H., Roth, F. P., & Schatschneider, C. (2004). Growth in early reading skills from kindergarten to third grade.*Contemporary Educational Psychology*, *29*(3), 312-332.

Spellings, M. (2005). Secretary Spellings announces growth model pilot [Press Release]. Washington, DC: U.S. Department of Education. Retrieved July 17, 2014 from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of Educational Psychology*, *86*, 221–234.

Stanford Achievement Test (9th ed.). (1996). San Antonio, TX: Harcourt Brace

Stevenson, C. E., Hickendorff, M., Resing, W., Heiser, W. J., & de Boeck, P. A. (2013).

Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence, 41*(3), 157-168.

Stevenson, D. D. (2004). What we know about correlates of reading. *Exceptional Children, 70,* 453–468.

Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development, 57*, 646–659. doi:10.2307/1130343

Stipek, D. (2002). At what age should children enter kindergarten? A question for policy makers and parents. Society for Research in Child Development Social Policy Report, 16, 1–16.

Storch, S., & Whitehurst, G. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*, 934–947. Retrieved from http://dx.doi.org/10.1037/0012-1649.38.6.934

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association, 81*(393), 82-86.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks. NCES 2009-004. *National Center for Education Statistics*.

Treiman, R. (2006). Knowledge about letters as a foundation for reading and spelling. *Handbook of orthography and literacy*, 581-599.

Treiman, R., & Broderick, V. (1998). What's in a name: Children's knowledge about the letters in their own names. *Journal of Experimental Child Psychology, 70,* 97–116.

Retrieved from http://dx.doi.org/10.1006/jecp.1998.2448

te Marvelde, J. M., Glas, C. A., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement, 66*(1), 5-34.

VanDerHeyden, A. M., Snyder, P. A., Broussard, C., & Ramsdell, K. (2008). Measuring response to early literacy intervention with preschoolers at risk. *Topics in Early Childhood Special Education*, *27*(4), 232-249.

van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. *New York*.

Verhelst, N. D., & Eggen, T. J. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. Arnhem: Cito.

Verkuilen, J. (2006). Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach by P. de Boeck and M. Wilson and Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models by A. Skrondal and S. Rabe-Hesketh. *Psychometrika*, *71*(2), 415-418.

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*(2), 318-336.

Vygotsky, L. S. (1962). Language and thought. *Massachusetts Institute of Technology Press, Ontario, Canada*.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., & Garon, T. (1997). Changing relations between phonological processing abilities

and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental psychology*, *33*(3), 468.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77-87.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60-62.

Wilson, M., & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing*, *28*(4), 441-462.

Wilson, M., Zheng, X., & McGuire, L. (2011). Formulating latent growth using an explanatory item response model approach. *Journal of applied measurement*, *13*(1), 1-22.

Wollack, J. A., Sung, H. J., & Kang, T. (2005, April). *Longitudinal effects of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The Impact of Compounding Item Parameter Drift on Ability Estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). ACER ConQuest: Generalized Item Response Modeling Software [Computer software and manual]. Melbourne, Victoria, Australia: Australian Council for Educational Research

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273-283). New York: Springer

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix: The glmer and bglmer code for LG-EIRM analysis

```
#ABC-M1

mod1<-glmer(resp ~ -1+shape+time+(1|itemNew)+(t1+t2+t3-1|person), data=abc,
family=binomial("logit"), control=glmerControl(optimizer="bobyqa", optCtrl =
list(maxfun = 100000)))


#ABC-M2

# drift effectof shape 1
s1dft2<-with(abc,factor(0+(time==2&shape==1)))
s1dft3<-with(abc,factor(0+(time==3&shape==1)))

mod1_s1<-glmer(resp~-1+shape+s1dft2+s1dft3+time+(1|itemNew)+(t1+t2+t3-1|person),
data=abc,family=binomial("logit"), control=glmerControl(optimizer="bobyqa", optCtrl =
list(maxfun = 100000)))


#ABC-M3

# drift effect of shape 2
s2dft2<-with(abc,factor(0+(time==2&shape==2)))
s2dft3<-with(abc,factor(0+(time==3&shape==2)))

mod1_s2<-glmer(resp~ -1+shape+s2dft2+s2dft3+time+(1|itemNew)+(t1+t2+t3-
1|person), data=abc,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun = 100000)))


#ABC-M4

# drift effect of shape 3
s3dft2<-with(abc,factor(0+(time==2&shape==3)))
s3dft3<-with(abc,factor(0+(time==3&shape==3)))

mod1_s3<-glmer(resp~ -1+shape+s3dft2+s3dft3+time+(1|itemNew)+(t1+t2+t3-
1|person),data=abc,family=binomial("logit"), control=glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun = 100000)))
```

```
#ABC-M5

# drift effect of shape 4
s4dft2<-with(abc,factor(0+(time==2&shape==4)))
s4dft3<-with(abc,factor(0+(time==3&shape==4)))

mod1_s4<-glmer(resp~ -1+shape+s4dft2+s4dft3+time+(1|itemNew)+(t1+t2+t3-
1|person), data=abc,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun = 100000)))


#ABC-M6

mod2_drift<-glmer(resp~ -1+shape*time+(1|itemNew)+(t1+t2+t3-1|person), data=abc,
family=binomial("logit"), control=glmerControl(optimizer="bobyqa", optCtrl =
list(maxfun = 100000)))


#ABC-M7

mod3_drift<-glmer (resp~1+time+shape+time:shape+prek+dis+ell+agefall+(1|itemNew)
+ (t1+t2+t3-1|person), data=abc, family=binomial("logit"),
control=glmerControl(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))


#ABC-M8

mod4_pr001<-bglmer (resp~1+time+shape+time:shape+prek+dis+ell+agefall +
time:prek+time:dis+time:ell+time:agefall+(1|itemNew) +(t1+t2+t3-1|person), data=abc,
family = binomial("logit"),control = glmerControl(optimizer="bobyqa", optCtrl =
list(maxfun = 200000)), cov.prior = wishart,fixef.prior = normal(cov=diag(2,24)))


#LS-M1

mod1<-glmer(resp ~ -1+sound+time+(1|itemNew)+(t1+t2+t3-1|person),
data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",check.conv.
grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M2

#drift effect of sound 1
s1dft2<-with(ls,factor(0+(time==2&sound==1)))
s1dft3<-with(ls,factor(0+(time==3&sound==1)))
```

```
mod1_s1<-glmer(resp~-1+sound+s1dft2+s1dft3+time+(1|itemNew)+(t1+t2+t3-1|person),
data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",check.conv.
grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M3

#drift effect of sound 2
s2dft2<-with(ls,factor(0+(time==2&sound==2)))
s2dft3<-with(ls,factor(0+(time==3&sound==2)))

mod1_s2<-glmer(resp~ -1+sound+s2dft2+s2dft3+time+(1|itemNew)+(t1+t2+t3-
1|person),data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",ch
eck.conv.grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M4

#drift effect of sound 3
s3dft2<-with(ls,factor(0+(time==2&sound==3)))
s3dft3<-with(ls,factor(0+(time==3&sound==3)))

mod1_s3<-glmer(resp~ -1+sound+s3dft2+s3dft3+time+(1|itemNew)+(t1+t2+t3-
1|person),data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",
check.conv.grad=.makeCC("warning",0.005),optCtrl = list(maxfun = 100000)))


#LS-M5

#drift effct of sound 4
s4dft2<-with(ls,factor(0+(time==2&sound==4)))
s4dft3<-with(ls,factor(0+(time==3&sound==4)))

mod1_s4<-glmer(resp~ -1+sound+s4dft2+s4dft3+time+(1|itemNew)+(t1+t2+t3-
1|person),data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",ch
eck.conv.grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M6

#drift effect of sound 5
s5dft2<-with(ls,factor(0+(time==2&sound==5)))
s5dft3<-with(ls,factor(0+(time==3&sound==5)))
```

```
mod1_s5<-glmer(resp~ -1+sound+s5dft2+s5dft3+time+(1|itemNew)+(t1+t2+t3-
1|person),data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",ch
eck.conv.grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M7

mod2_drift<-glmer(resp ~ -1+sound*time+(1|itemNew)+(t1+t2+t3-1|person),
data=ls,family=binomial("logit"),control=glmerControl(optimizer="bobyqa",check.conv.
grad=.makeCC("warning",0.005), optCtrl = list(maxfun = 100000)))


#LS-M8

mod3<-glmer(resp ~ 1+time+sound+time:sound+prek+dis+ell+agefall+(1|itemNew)+
(t1+t2+t3-1|person), data=ls,family=binomial("logit"),
control=glmerControl(optimizer="bobyqa",check.conv.grad=.makeCC("warning",0.005),
optCtrl = list(maxfun = 200000)))


#LS-M9

mod4_pr<-bglmer(resp~1+time+sound+time:sound+prek+dis+ell+agefall+time:prek+
time:dis+ time:ell+time:agefall+(1|itemNew)+(t1+t2+t3-1|person),data=ls, family =
binomial("logit"),control = glmerControl(optimizer="bobyqa",
check.conv.grad=.makeCC("warning",0.005),optCtrl = list(maxfun = 100000)), cov.prior
= wishart, fixef.prior = normal(cov=diag(2,27)))
```