Title
Practical Formulations of the Latent Growth Item Response Model

Author
McGuire, Leah Walker

Publication Date
2010

Peer reviewed|Thesis/dissertation

Practical Formulations of the Latent Growth Item Response Model

by

Leah Walker McGuire

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Education
in the
Graduate Division
of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Alan Hubbard
Professor Frank Worrell

Spring 2010

Abstract

Practical Formulations of the Latent Growth Item Response Model

by

Leah Walker McGuire

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

Growth modeling using longitudinal data seems to be a promising direction for improving the methodology associated with the accountability movement. Longitudinal modeling requires that the measurements of ability are comparable over time and on the same scale. One way to create the vertical scale is through concurrent estimation with identification of groups (Bock & Zimowski, 1997). However, there are concerns about how well this vertical scale will function using longitudinal data with few common items between years (Briggs et al., 2008). Other concerns about the adequacy of this and other vertical scaling strategies arise when the common items shift over time. This study explores these two practical issues in application of a Latent Growth Item Response model (LG-IRM). To illustrate how psychological constructs could be tracked over time, an application to psychological data is also included.

Since the number of common items between years can be few, it is important to ensure that the growth modeling procedure can produce good estimations in this situation. A concurrent estimation procedure of the scales is examined. To verify the estimability of the growth model using concurrent estimation, a simulation study is conducted. The LG-IRM is then applied to real data from the LSAY. A study of item shifts over time is also included. Models that do not consider item shifts are compared to those that do. An extended version of the Latent Growth Item Response Model is proposed in which estimates growth parameters are produced while allowing for item shift over time.

Item Response Theory (IRT) has been applied extensively to research in education. Its use in modeling achievement and student ability has been demonstrated using various IRT model formulations. To a lesser degree IRT has also been applied to personality research. Although the conceptualization of the domain, the item types, and the measurement goals are often different in personality research, IRT tools can still provide valuable information. The LG-IRM also provides a way to explore questions in personality research. For instance, it can provide information on the stability of self-esteem in the Black population over time. In this study the Latent Growth Item Response Model is extended to include polytomous item types. The LG-IRM is then applied to data from a longitudinal study of Black Self-Esteem.

Keywords: Growth Modeling, Item Response Theory, Latent-Growth Item Response Model, Longitudinal Survey of American Youth, Measurement Invariance, National Educational Longitudinal Survey, National Survey of Black Americans, Self-Esteem, Vertical Scaling.

Dedication

To my husband and to my sisters for making it a merry way.

Table of Contents

List of Figures

List of Tables

Symbols

Item Difficulties and Steps

δ_i	difficulty of item *i*
$\boldsymbol{\xi}$	a vector of item difficulties and step parameters
τ_{ik}	k^{th} step parameter for item *i.*

Latent Ability

θ_{pt}	ability of person *p* at time *t*
θ_{pb}	ability of person *p* at baseline, subscript b pertains to baseline
θ_{pg}	linear growth in ability of person *p,* subscript g pertains to growth
$\bar{\theta}_{pt}$	mean ability at time t
$\boldsymbol{\vartheta}_p$	vector of latent abilities

Responses

X_{ipt}	response of person *p* at time *t* to item *i*

Sample

N_p	sample size (number of persons)
N_{pt}	sample size at time t

Model Specification

A	design matrix
a_{ij}	design vector
B	scoring matrix
b_i	scoring vector

Results

SE	standard error
SD	standard deviation
CI	confidence interval
T	standardized residual
t	t-statistic used in t-test

Introduction

One stipulation of The No Child Left Behind Act (NCLB, 2002) is that States implement valid accountability systems to measure progress. In many cases, the measurement of adequate yearly progress (AYP) is conducted by comparing scores of successive cohorts of students. Many studies do not employ repeated measures designs due to a lack of time or resources to take multiple measurements. However, states can apply for multi-million dollar grants to design and implement longitudinal data systems under the Educational Technical Assistance Act (NCLB, 2002). Since repeated measures designs present a considerable strain on fiscal resources in any research study, it is important to ensure that the measurements being taken are suitable for comparison over time and that they are valid for diverse groups of students. This dissertation presents practical formulations of an item response model for change over time.

Three example data sets are used to illustrate practical application of the model. The example data, taken from publicly available longitudinal studies, provide the basis for exploration of the model in Chapters One through Three. Chapter One also includes a simulation study to verify the estimability of the model. In each study, emphasis is also given to the presentation of the results.

In Chapter Two, the Latent Growth Item Response Model LG-IRM is applied to the mathematics sections of the Longitudinal Survey of American Youth (LSAY). The mathematics assessment included different sets of forms in each year. In order to estimate the LG-IRM using these data, the issues related to equating and vertical scaling will be explored. One way to create the vertical scale is through concurrent estimation with the identification of groups (Bock & Zimowski, 1997). However, there are concerns about how well this vertical scale will function using longitudinal data with only a few common items between years (Briggs, Weeks, & Wiley, 2008). A simulation study will be conducted to verify that the concurrent estimation of the vertical scale is appropriate when modeling growth using the LG-IRM. The results of the simulation study will illustrate whether it is appropriate to use concurrent estimation of the LG-IRM parameters when using vertically scaled data from an achievement test. Once the use of the procedure is verified, the LG-IRM can then be applied to actual data from the LSAY.

The LG-IRM can also be an effective tool for personality research. In Chapter Three, the LG-IRM is applied to a self-esteem measure. Although the conceptualization of the domain, the item types, and the measurement goals are often different in personality research, IRT tools can still provide valuable information. This will be demonstrated using longitudinal measurements of self-esteem from the National Survey of Black Americans (NSBA). In this study, the LG-IRM must also be extended to include polytomous item types, since the self-esteem measure includes Likert-style items. The results from the application of Partial Credit formulation of the LG-IRM are used to illustrate the possibilities for application of the model in personality research.

In Chapter Three, issues of item shift are explored using data from the National Educational Longitudinal Survey (NELS). This chapter expands on some points discussed in Chapter Two. Specifically, the scaling of forms across years and the NELS 1988 History subtest (NELS, 2000) show that items do indeed shift over time. Although such shifts in history can be explained using arguments about the changing curriculum and lack of a vertically articulated developmental scale, this example is used to illustrate how the LG-IRM can be formulated to handle these shifts in other domains as well. This study invokes the frameworks of differential item function (DIF) to explore not only the more traditional group-wise DIF, but also to explore item shifts over time.

Acknowledgements

My heartfelt thanks to my committee for their detailed feedback and advice. Thanks to the BEAR Seminar audiences for providing helpful feedback on this work. Finally, thanks goes to the Models of Assessment research group for support both inside and outside the classroom.

Chapter 1: Issues in Application of the Latent Growth Item Response Model for Achievement Data

Issues in Application of the Latent Growth Item Response Model for Achievement Data

With the advent of No Child Left Behind, it has become increasingly important to track gains in student outcomes (NCLB, 2002). Mandated reporting of adequate yearly progress (AYP) requires that states report on their progress toward ensuring that all students reach minimal levels of proficiency. However, as discussed in the introduction, it has been difficult to determine how progress should be estimated and reported in the complex context of educational measurement. Current questions center on if growth should be quantified as gain scores or in other metrics, and on how to create these growth metrics. This paper seeks to join the conversation by providing a way to calculate growth. Answering the call of the National Research Council and National Academy of Education (NRC & NAE, 2010) to provide more evaluative studies of growth metrics, this research employs a simulation study to verify that feasibility of the application proposed. In doing so, this study seeks to promote a method for estimating growth that can be used to provide meaningful estimates of student learning in complex educational contexts.

A simple way of reporting progress is to calculate gain scores. A gain score can be as simple as the difference between test scores from one year to the next. This method is attractive for use at the policy level because of the simple logic involved (Rogosa & Willett, 1983). However, gain scores suffer from the criticism that only two points in time are used. As described by the NRC and NAE (2010) joint report, a gain score based on observations of only two points in time may be more attributable to measurement error than actual growth. Also, to isolate learning attributable to a certain course or period of time, it may be necessary to control for background variables as is often done in value-added modeling (VAM) (Ballou, Sanders, & Wright, 2004).

VAM is a promising way to answer questions about how student characteristics and backgrounds influence student learning. An *accountability coefficient* is used to describe model parameters that are included to show the influence of a teacher or school on student learning. For instance, a VAM might include an accountability coefficient for a teacher, a school administrator, or a parent. A positive coefficient would show that the party is having a positive influence on student learning whereas a negative coefficient would show the opposite. Because of the many possible applications of accountability coefficients, including merit pay and sanctions, it is becoming increasingly important to create quality growth models for use in the educational context.

One of the major areas where attention is needed in order to create VAM is in defining the appropriate measurements used for growth modeling. The measurements are selected in a separate step, before the background variables and accountability coefficients are added. This measurement process involves collecting observations at multiple points in time and vertically scaling the scores, so that all of the measurements are on the same scale. The methods available for vertically scaling these data are within the common-item non-equivalent group family of linking methods (Kolen & Brennan, 1995). The method must be chosen carefully given that the size of growth can differ based on how the scale was constructed (Briggs & Weeks, 2008; Martineau, 2006; Tong & Kolen, 2007). Within IRT equating, these methods include concurrent estimation, item characteristic curve methods such as the Stocking-Lord (Stocking & Lord, 1983) or Haebara (1980) approaches, and mean equating methods (Marco, 1977).

Until recently, few studies have used actual longitudinal data to create the vertical scale (Briggs et. al., 2008). Instead, vertical scales were created using data from different grade levels, using common items between the tests given to the different grade levels. Therefore, it is important to provide evidence that the chosen vertical scaling method works well for

longitudinal data, since vertical scaling can be such an important and influential step in VAM (Briggs & Weeks, 2008; Martineau, 2006). However, longitudinal designs can be expected to have much more overlap between the populations than do most equating studies, because it is a large cost to a testing program to give multiple forms to the same student. The results of these equating methods may differ if longitudinal data are used, but since few studies actually use longitudinal data, it is difficult to know what these differences may be. Therefore, it is important to ensure that any procedure used as part of creation of the growth model works well with longitudinal data.

Currently, considerable debate revolves around the best method for constructing the vertical scale as well (Briggs and Weeks, 2008; Martineau, 2006). Concurrent estimation uses one run to estimate a single set of item difficulty parameters for the entire bank of items used in all of the test forms. The separate linking methods use separate runs of each test form, which are then linked together. Although a complete review of this debate is beyond the scope of this paper, it is still important to justify the use of any particular method in this context. Concurrent estimation was chosen for this application for several reasons. First, it has been noted that concurrent estimation may be sensitive to the number of common items (Briggs & Weeks, 2008). All previous studies of the LG-IRM have been based on samples with large numbers of common items across all occasions and have been conducted using concurrent estimation (Wilson, Zheng, & Walker, 2007). As this paper seeks to build on the previous work on the LG-IRM, it was logical to begin with the same estimation procedure to determine if concurrent estimation can be used with more realistic data.

It is important to note that the performance of concurrent estimation has also been found to be sensitive to the fit of the IRT model. When the IRT model fits well, even with few common items, it has been shown that concurrent estimation performs better than the characteristic-curve methods (Kim & Cohen, 1980). In the case of a simulation, which is built from the model, concurrent estimation should not pose problems, as were found in another simulation study that compared concurrent estimation to separate linking methods (Hanson & Beguin, 2002). The example data were also examined for model fit to ensure that the lack of model fit, which has been shown to produce different results by linking method (Beguin, Hanson, & Glas, 2000), is not introducing error.

It has also been suggested that concurrent estimation may be a less popular method for use in testing agencies because their pools of items are typically quite large, and it can be difficult to estimate so many parameters in a single run (Briggs et al., 2008). However, the simulation and example have moderately-sized item pools accompanied by large sample sizes, so estimation of all of the item parameters was not thought to be a problem in this study. Finally, when conducting a simulation, feasibility and labor costs must be taken into account. Since multiple packages of software must be used, the assembly and vertical scaling of each data set requires multiple imports, exports, and manipulations. Concurrent estimation is the least labor-intensive method. Therefore, the simulation and example in this study approach the issue of vertical scaling by using concurrent estimation.

Another area of concern in creating the vertical scale is in how the choices made in creating the scale affect score interpretation. For instance, because common items are required for creation of the vertical scale, it is tempting to overload the forms with too many common items. This results in a test that measures a construct common to learning at all stages, but not one of development over time (Schmidt, Hoang, & McKnight, 2005). The simulation shows how the baseline ability and growth can be calculated on the same scale as the set of items. If the items

are designed to match certain levels of the developmental scale, then this allows for interpretation of growth in terms of the developmental scale. This result, combined with the Wright Maps and interpretive figures used for description of the example data, shows how incorporation of the growth model into the item response model can attend to important concerns in growth modeling.

Wright Maps are used to illustrate the connection between student gains and the distribution of the items. They also illuminate another important point in applying the results from growth models to VAM, and clarify that the growth being estimated is measured in the same terms as the test. This connection is taken for granted by many in the educational community involved in making measurement decisions. However, the far-reaching audience of stakeholders who may be consumers of VAM estimates may generalize the results from one assessment to a broader set of skills than was actually measured. Thus, it is important to make it clear that the growth reported is in terms of the range of abilities measured by the items given (Lockwood, Hamilton, McCaffrey, Stecher, & Martinez, 2007), which is made possible by the incorporation of the growth model into the item response model.

The Present Study

In this study, I present a full simulation of vertically scaled data that was designed to mimic realistic data. I discuss how the design matrix can be manipulated to match the design of the vertical scale. I also show how well the estimated parameters match the generated parameters for this simulation example. Finally, to show how the LG-IRM would work in a practical situation, data from three waves of mathematics assessment are used. Various representations of the person and item estimates are produced as examples of what could be produced for applications like VAM.

Method

Simulation.

Values for the generating parameters were chosen to mimic a mathematics test given over three years. Personal abilities and growth values were generated to be consistent with findings about learning in mathematics in Grades 10 through 12, which show that only small positive growth is achieved in these years (Ralph & Crouse, 1997; Rampey, Dion, & Donahue, 2009). The ability and growth were designed to be correlated positively to match some of the findings of mathematics growth by school using the LSAY (Seltzer, Choi, & Thum, 2002). The participation of simulated respondents was also considered. If it were true that all students continue to participate in the same assessment programs over time, then this consideration would not be made. However, it is well documented that sample sizes tend to decrease over time as participants or students drop out (Little, 1995). The data were designed so that an increasing proportion of students were randomly missing from each year of testing. The amount of missing data was set at 5%, 12% and 20% in Years 0, 1, and 2 respectively. Since the data were created through a random draw from the uniform distribution, the patterns of missing data were truly missing at random (MAR) (Rubin, 1973) and did not merit as much worry as non-random patterns of missing data, such as high school drop out.

Forms increased in difficulty to match tests to design increasing levels of ability. The parameters of item difficulty were generated and grouped so that the form in each year had increasing average difficulty. The grouping of the items in each form was also considered for

vertical scaling purposes. A vertical scale combines links between forms from year to year, and these links, together, create a scale over a span of years. The items used to create these links are referred to as *common items*. In the simplest case, the same core set of common items would be present on the test in each year. However, in practical situations such as state-mandated testing, it is very difficult to maintain a core set of items on the test every year, given that few state testing programs allow large numbers of out-of-level items (Roeber, Bond & Connealy, 1998). Thus, it is more likely that the common items that do exist serve as links only between consecutive years. The simulated data used in this study were designed to match practical settings by including small but positive simulated growth, test forms of increasing difficulty, and links only between consecutive years.

Some decisions also have to be made about constructing the simulated data sets. Data from one-dimensional item response models can be generated in many item response modeling software packages. ConQuest (Wu, Adams, & Wilson, 1998) specifically simulates data from the Rasch family of models. In addition, data from the multidimensional within-item models of the Rasch family can be simulated in ConQuest. Currently, no item response modeling software packages have an embedded routine to generate data from the multidimensional within-item models of the Rasch family, such as the LG-IRM. Hence additional code or data manipulation is needed to produce simulated data from the LG-IRM. The method of generating data from the within-item multidimensional model for this study is described briefly below. Additional information, including the code used for this example, is included in Appendix A.

First, correlated personal abilities and the complete set of item difficulties were generated. The baseline and linear growth ability parameters can be drawn from a bivariate normal distribution. The generations of item difficulty can be performed in almost any statistical program by considering these values to be random draws from a specified distribution. The generations can also be performed in many item response modeling software packages. Finally, the simulated responses can be generated using the probabilistic relationship defined by the LG-IRM. In the linear growth specification of the LG-IRM (Equation 1), the effective personal ability in each year is a linear combination of the baseline ability and growth.

Simulation procedure.

The aim of performing a simulation is to test for any systematic bias or distortion in the estimation. Although any individual simulation result may show error, if this error averages to a very small amount over a number of repetitions, then there is no evidence of significant bias. Twenty simulated data sets were created for use in the parameter recovery study. In order to create these data sets, the item difficulties and person abilities were simulated. The item difficulties were simulated once. This set of item difficulties is used across all simulation replications. 20 different sets of person abilities were simulated according to the design. These item responses were generated using these abilities and item difficulties. The parameters of the LG-IRM were then estimated using ConQuest. The personal abilities were estimated using maximum-likelihood (MLE), weighted-likelihood (WLE) (Warm, 1989), and Estimates a Posteriori (EAP) (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). EAP was found to converge faster with the simulated data, possibly because there was a significant amount of missing data due to the design and concurrent estimation procedures. However likelihood-based approaches have been used and found to effective in several other studies of similar models (Cagnone, Moustaki, and Vasdekis, 2009; te Marvelde & Glas, 2006; Wilson et al., 2007). So the WLEs are used for presentation in this study in order to maintain consistency with these studies. Future studies may look at systematic differences among types of estimates. Finally, the

estimates were compared to their generating values for evidence of bias and overall parameter recovery.

The model.

The LG-IRM is a unique model because it is formulated and estimated as an item response model. As described in the notation section, the subscript *p* is used to index the persons and the subscript t is used to denote points in time. In addition to the time subscript, *t*, subscripts *b* and *g* are used to reference baseline and linear growth. The linear growth version of the LG-IRM defines a person's ability at any point in time, θ_{pt}, as a function of the person's ability a baseline ($t = 0$) and future points in time ($t = 1, 2, \ldots$) using a intercept (θ_{pb}) and a slope term (θ_{pg}).

$$\theta_{pt} = \theta_{pb} + t\theta_{pg} \tag{1}$$

Since a linear model is specified, the ability at each individual time point does not need to be estimated. Only the baseline ability and growth need to be estimated. The LG-IRM formulation (Wilson et al., 2007) then shows how the conditional probability of a correct response from person j to any item at time t (X_{ipt}) given the person ability at that time point, θ_{pt}, can be written as a function of the difference between person p's ability at time t (θ_{pt}) and the difficulty of item i (δ_i).

$$\mathrm{P}(X_{ipt} = 1 \middle| \theta_{pt}) = \mathrm{P}_{ijk} = \frac{\exp(\theta_{pt} - \delta_i)}{1 + \exp(\theta_{pt} - \delta_i)} \tag{2}$$

The equations presented here communicate clearly the type of growth defined by the application of the LG-IRM in this study. However, a few other practical features deserve mentioning. One feature that must be specified by the user is the *scoring matrix*. The *scoring matrix* is used to define the mapping of items onto each of the dimensions. For instance, an item that is only present in the first wave of testing, should only map onto the baseline ability dimension whereas items present in the second wave would map onto both the baseline ability dimension and the linear growth dimension. Another feature, called the *design matrix*, is used to set the proper constraints for scaling and estimation. To satisfy the scaling considerations, the design matrix is used to hold the difficulties of the common items to be constant over time. To ensure that the model is identified, the design matrix is used to constrain the mean of the item difficulties to zero. Specification and examples of the scoring matrix and design matrix is discussed in further detail in the sections that follow.

The LG-IRM accomplishes several goals in one run of ConQuest. It puts all of the responses from the different forms given in each year, estimates the item difficulties, and produces the personal ability and growth. Although this is a convenient and powerful way to estimate the growth model, it is important to note the implicit choices that are made to estimate the model in one run. First, in order to put all of the forms onto the same scale, ConQuest employs an implicit *vertical scaling* procedure called concurrent calibration. Concurrent calibration is just one way to create a vertical scale. For instance, some procedures use the estimates from separate calibrations to create the vertical scale. Based on the research cited in earlier sections, it is thought that

concurrent estimation will work well for the simulation and example in this study. It is important to note that this choice.

Simulated item difficulties.

A set of 165 item difficulties was generated to serve as the bank from which forms of varying difficulty could be assembled. The item difficulties values are included in a four-logit interval, which might be a reasonably large enough interval to track student progress across school grades. A smaller interval might be more appropriate for a simulation designed to mimic a targeted assessment for students within a narrow developmental range. The code used to draw the item difficulties from a uniform (-2, 2) distribution is shown in Appendix A. The three test forms were designed specifically to have different levels of difficulty and different numbers of items. The average difficulty of the forms increases each year. The average difficulties of each form are shown in Table 1. The Wave 0 form might mimic a mathematics test given in the early middle school grades. It has a low average difficulty (-0.3669 logits) on the overall scale. It also contains the most items (75). A longer test form could match a real situation in which longer, exit examinations are given as a student moves from middle to high school or graduates from high school. The Wave 1 form has increased average difficulty (0.128 logits) compared to the Wave 0 form. It contains 65 items, 15 of which are common to the Wave 0 form and 15 of which are common to the Wave 2 form. These sets of common items do not overlap with each other; the link between Wave 0 and Wave 2 can be established only through using the results from the Wave 1 test. The average difficulty of the Wave 2 test is 0.5042 logits. It contains 55 items, 15 of which are common to the Wave 1 test.

Table 1: Characteristics of Vertically Scaled Test Forms for Full Simulation

Form	Average Difficulty	Number of Items	Number of Common Items
Wave 0	-0.3669	75	-,15
Wave 1	0.1283	65	15, 15
Wave 2	0.5042	55	15,-

Note. The same set of item difficulties were used for all of the simulations. The number of items common with consecutive wave shows the number of items common to the previous wave and following wave separated by a comma. The average difficulty is calculated as an average of the difficulties over the items in each form.

Examples of simulated data are shown in Figure 1. These excerpts show the pattern of responses for three students. The first example is a simple case of vertically scaled data. The second example is an excerpt from the actual data used in this study. In both examples, the left-most column contains the student identification number. This is an important feature of any longitudinal analysis because it allows for the matching of the same student scores from one year to the next. The next column shows the year of administration. These years should start with zero to match with the parameterization of the LG-IRM in ConQuest. The data should be ordered by student and then year so that ConQuest's input format will recognize the observations correctly. Finally, the item responses are shown in the remaining columns. Each item occupies one column.

Items that are common to multiple years will show non-missing responses in multiple years. In this first example, the set of common items across all years shows non-missing responses, or entries in the same column, for Years 0, 1, and 2. Items that were not present on a certain test form receive missing responses. These missingness-by-design patterns are modeled using the design matrix. The staircase pattern, shown in the first example of Figure 1, was created visually

by ordering the items by difficulty. Since the forms were designed to increase in difficulty each year, the staircase pattern naturally shifts to the right. In practical settings, the data organized by year may not look as neat. The actual data used for the simulation is shown in the second example of Figure 1. This example differs from the first in that the items have not been arranged by difficulty, some personal data are missing, and common items exist only between consecutive years. In these cases, the specification of the design matrix will require careful mapping. An example of a real design matrix is explained in the example taken from the LSAY mathematics test.

Figure 1. Examples of simulated data using a two different designs; one with the same common items across six waves of testing, and one with common items between consecutive years of testing only for three waves of testing.

Example 1: Common Items Link Six Waves of Testing.

person	year	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	i24	i25	i26	i27	i28	i29	i30	i
1001	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.	.	.	.
1001	1	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.
1001	2	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1001	3	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1001	4	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1001	5	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1002	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	.	.	.	.	.	.	.	.	.	.	.
1002	1	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.
1002	2	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1002	3	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1002	4	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1002	5	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1003	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	.	.	.	.	.	.	.	.	.	.	.
1003	1	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.
1003	2	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1003	3	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
1003	4	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1003	5	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Example 2: Common Items Between Consecutive Waves of Testing Link Three Waves.

person	year	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	…	i35	i36	i37	i38	i39	i40	i41	i42	i43	i44	…	i70	i71	i72	i73	i74	i75
10001	0	0	0	0	0	0	0	1	0	1	1	0	1	1	0	1	0	1	0	0	0	0																	
10001	1																1	0	1	1	0	1	0	0	0	1	1	0	1	1	0	1							
10001	2																						1	1	1	1	0	1	0	1	0	1		1	1	1	0	0	1	0
10002	0	1	0	1	1	1	1	0	0	0	0	1	0	1	0	0	0	0	1	1	0																		
10002	1																1	0	1	0	0	1	1	1	1	1	1	1	0	1	0	1							
10002	2																						0	1	0	0	1	1	1	0	1	0		1	1	0	0	0	0	0
10003	0	1	1	0	1	0	1	1	0	0	0	1	0	1	1	1	1	1	0	1	0																		
10003	1																1	0	1	0	0	1	1	1	0	1	0	1	1	1	1	1							
10003	2																						1	1	0	0	1	0	0	1	0	1		1	1	0	1	0	0	1

Simulated abilities.

For each of the 20 replications, a different set of person ability and growth values was generated. The total number of cases for each replication, N_p, is held constant at 2,200 cases. The number of persons present in the baseline year (N_{po}) is roughly 95 per cent of the total number of cases used in each analysis. N_{po} ranges from 2,075 to 2,113. The sample mean ($\bar{\theta}_{p0}$) was also calculated to show the average initial ability for each simulation. The initial ability ranges from -0.532 logits to -0.474 logits. These values were calculated over the number of persons present in Year 1. The number of persons was designed to decrease in each following year. The number of persons in Year 1 is roughly 88 per cent of the total number of cases and the number of persons in Year 2 is roughly 80 per cent of the cases. The effective abilities at Time 1 and Time 2 were calculated by applying the formulas in Equation 2, using the simulated growth values. The average effective abilities at each time point are also shown in Table 23. They are both calculated as averages over the number of persons present in each year. The average effective ability in Year 1 ($\bar{\theta}_{p1}$) ranges from -0.243 to -0.162 logits. The average effective ability in Year 2 ($\bar{\theta}_{p2}$) is higher, with values between 0.033 and 0.175 logits.

Simulated responses.

The simulated item difficulties and effective personal abilities were used to generate a separate data set for each year. These data sets were then combined by matching the student number and item. The resulting data were similar to those shown in the second example of Figure 1. Next, to ensure that the data had been generated properly according to the LG-IRM specification in ConQuest, the data were used to estimate the LG-IRM parameters. The goal of this preliminary estimation is to produce the generating parameters to be used for the simulation study. Other simulations might rely on the estimated parameters from real data to ensure that the simulation is properly designed, but real data of this type were not available for this study. This step also ensures that the constraints placed in the data generation step match those of the estimation, particularly in reference to the item distribution and common item linkages. Next, a new data set of responses was produced using the person ability and growth estimates as well as the item difficulties. Although students who only participated in one wave beyond the baseline would receive a growth value, attention was paid to the original data participation designs. Thus, responses were generated for each student by year according to the designed participation pattern.

Table 23*:* Simulated Sample Participation

Replication	N_p	N_{p0}	N_{p1}	N_{p2}	$\overline{\theta}_{p0}$	$\overline{\theta}_{p1}$	$\overline{\theta}_{p2}$
1	2,200	2,077	1,711	1,393	-0.510	-0.231	0.047
2	2,200	2,075	1,717	1,378	-0.524	-0.243	0.103
3	2,200	2,082	1,701	1,371	-0.518	-0.232	0.033
4	2,200	2,101	1,703	1,434	-0.479	-0.149	0.175
5	2,200	2,090	1,705	1,385	-0.515	-0.224	0.053
6	2,200	2,102	1,729	1,434	-0.499	-0.199	0.094
7	2,200	2,083	1,689	1,418	-0.518	-0.209	0.153
8	2,200	2,079	1,717	1,394	-0.492	-0.237	0.047
9	2,200	2,094	1,696	1,424	-0.491	-0.162	0.158
10	2,200	2,105	1,709	1,426	-0.526	-0.217	0.060
11	2,200	2,098	1,689	1,429	-0.489	-0.186	0.153
12	2,200	2,087	1,683	1,392	-0.476	-0.181	0.102
13	2,200	2,088	1,730	1,407	-0.496	-0.184	0.135
14	2,200	2,096	1,728	1,446	-0.510	-0.195	0.122
15	2,200	2,103	1,672	1,463	-0.498	-0.193	0.107
16	2,200	2,088	1,672	1,363	-0.505	-0.228	0.061
17	2,200	2,080	1,682	1,412	-0.474	-0.172	0.086
18	2,200	2,080	1,726	1,395	-0.478	-0.159	0.109
19	2,200	2,113	1,723	1,401	-0.532	-0.233	0.053
20	2,200	2,103	1,922	1,718	-0.478	-0.191	0.100

Note. Missingness processes were designed to be at random, but with increasing missingness proportions in each wave. The sample mean in each wave taken is calculated as the average of the effective personal abilities in each wave for the persons present in each wave of testing.

Simulation results.

Three parameter recovery values were calculated to evaluate the results of the simulation. These are the Pearson correlation, the average signed bias (ASB) and the root mean-square error (RMSE). These were calculated for the item difficulties (Table 3) and the personal abilities and growth (Table 4). The Pearson correlation is calculated to show how closely associated the estimates are with their generating values. The ASB is an index of direction bias. If it is positive, then the estimates tend to be larger than the generative values. If it is negative, the estimates tend to be smaller than the generating values. Any consistent pattern of signs in the ASB across replications signals that the estimation is producing biased results in one direction. The RMSE is an unsigned measure of parameter recovery. The size of the RMSE is an indication of how close the estimates are to the generating parameters. The ASB and RMSE are in logits, and provide a measure of parameter recover in the same units as the item difficulties and person abilities. For this example, the EAP estimates were used for parameter recovery. For the amount of missing data in this simulation example, EAP estimates resulted in convergence for a larger number of students than did the likelihood-based approaches. This is to be expected because the EAP estimation procedure can supplement the estimation with information from the prior distribution.

Table 3: Item Parameter Recovery

Replication	Pearson Correlation	ASB	RMSE
1	0.9656	0.00093	0.02027
2	0.9658	-0.00077	0.02022
3	0.9658	0.00023	0.02022
4	0.972	-0.00094	0.01848
5	0.9819	-0.00147	0.01477
6	0.9766	-0.0015	0.01688
7	0.9657	-0.00183	0.02024
8	0.9702	0.00115	0.01888
9	0.9775	0.00129	0.01657
10	0.9775	0.00131	0.01642
11	0.9846	0.00139	0.01367
12	0.9829	0.0014	0.01442
13	0.9928	-0.00116	0.00936
14	0.9769	-0.00173	0.01663
15	0.983	0.00081	0.0143
16	0.9948	0.0008	0.0081
17	0.9721	-0.00159	0.01835
18	0.9657	-0.00183	0.02024
19	0.9802	-0.00145	0.0001
20	0.9998	-0.00109	0.01553
Average	0.9776	-0.00030	0.01568

Table 4: Personal Ability and Growth Recovery

Replication	Pearson Correlation	$N_{\theta b}$	$ASB_{\theta b}$	$RMSE_{\theta b}$	$N_{\theta g}$	$ASB_{\theta g}$	$RMSE_{\theta g}$
1	0.96018	2077	-0.01318	0.54825	2026	0.00765	0.36969
2	0.96281	2075	0.09179	0.53724	2032	-0.08966	0.46413
3	0.95899	2082	0.04683	0.54854	2030	0.04177	0.42839
4	0.95688	2101	-0.00214	0.56432	2049	0.00189	0.43597
5	0.95574	2090	0.07139	0.58012	2033	-0.06966	0.34506
6	0.96285	2102	0.09473	0.52931	2051	-0.09341	0.45810
7	0.95597	2083	0.09184	0.56866	2028	-0.08459	0.34210
8	0.95506	2079	-0.01907	0.59392	2029	0.00993	0.44894
9	0.96560	2094	-0.02745	0.49646	2042	0.02389	0.46368
10	0.95514	2105	0.04302	0.58087	2050	0.03671	0.30925
11	0.95402	2098	-0.02352	0.59437	2039	0.02184	0.38724
12	0.96430	2087	0.07575	0.50634	2032	-0.06692	0.44753
13	0.96193	2088	0.08703	0.54267	2019	-0.08284	0.46297
14	0.95393	2096	0.08662	0.60762	2045	-0.08590	0.44063
15	0.96353	2103	0.03911	0.50739	2046	0.03555	0.40945
16	0.96035	2088	0.03682	0.54270	2037	0.02747	0.40508
17	0.95781	2080	-0.02581	0.56026	2030	0.02202	0.48776
18	0.97607	2080	0.07487	0.46105	2029	-0.07239	0.39518
19	0.95591	2113	0.07322	0.57020	2061	-0.07236	0.42124
20	0.94886	2103	-0.03403	0.62108	2052	0.03266	0.44085
Average	0.95930	2091	0.03839	0.55307	2038	-0.02282	0.41816

Each parameter recovery table (Tables 3 and 4) shows the list of indices for each replication. The Pearson correlation for the item parameters is high. Values ranged from 0.9656 to 0.9998. The average value for the set of 20 replications was 0.97757. In other simulation studies of the LG-IRM, it was found that the item estimates were consistently underestimated (Wilson et al., 2007). In other words, the ASB values for the item estimates were consistently below zero. However, this is not the case with this example. It is possible that the underestimation could have been a default constraint set by the ConQuest program. The default constraint is set on the last item of the set. Thus, the last item acts as a balancing item. If the item difficulties are generated without regard to this constraint, or not arranged with this constraint in mind, then there could be a mismatch between the generating parameters and the estimation constraints. Since the last item is used as a balancing item for the entire distribution of item difficulties, the rest of the difficulties could be underestimated.

The data for this simulation were constructed using parameters produced by the ConQuest estimation of the LG-IRM. Thus, the difficulties were already arranged in accordance with the estimation constraints. It is, therefore, not surprising to see the random pattern of signs of the ASB values in Table 4. The set of replications does not show consistent over- or under-estimation of the item difficulties. The ASB values for the items (Table 3) are small. They range from -0.00183 to 0.0014 logits with an average of -0.0003 logits. These values are the smallest of the three sets of parameter recovery metrics, probably because positive and negative values can cancel each other out, resulting in an average closer to zero. The RMSE values are also small, with values between 0.0001 and 0.02027 logits. The results in Table 3 show a good estimation of the item difficulties.

The results for the personal ability and growth are well-recovered overall. The Pearson correlation values range from 0.94886 to 0.97607, with an average of 0.95930. If these values are considered to be slightly low for parameter recovery studies, it is helpful to know that the size of the ASB and RMSE values, are reasonably small. These indices were calculated for the baseline ability ($ASB_{\theta 1}$ and $RMSE_{\theta 1}$) and growth ($ASB_{\theta 2}$ and $RMSE_{\theta 2}$) separately, as shown in Table 4.

The baseline ability represents the first dimension in the LG-IRM. It is important to correctly target the baseline ability in a growth study so that change can be measured in reference to the correct starting point. The simulation results in Table 4 show that the LG-IRM can locate the initial ability well. The ASBs for the baseline ability range from -0.03403 to 0.09473 logits. The RMSEs show values between 0.46105 and 0.62108 logits. These values are small considering that the span of the items was designed to include a four-logit range. In addition, it should be considered that some of these values were calculated for students who participated only in the baseline administration, as well as students who might have only participated in one additional year of testing. Due to the missingness patterns, EAP estimation was chosen. In comparison to likelihood-based approaches, it is known that the EAP estimates will be larger (Briggs & Weeks, 2008). This effect has been cited as a reason that the consideration of vertical scale construction is critical when proposing VAM. It is, therefore, quite possible that the choice of EAP estimation and the participation patterns of students could have affected the overall parameter recovery. In spite of this possibility, the simulation results show that the LG-IRM is still able to recover the initial values well, even with some missing data.

The growth represents the second dimension in the LG-IRM. If growth relative to a starting point is to be considered as a measure of teacher quality, then it is important to calculate the amount of growth correctly. The ASB values for growth range from -0.09341 to -0.04177. The

RMSEs are near 0.40 logits. This distance on the scale could represent the difference between students being classified within a relatively narrow range of abilities. Based on the average growth from year to year, this would probably result in the classification of students within the same grade-level if a developmental scale were attached to the simulation design. In the real data example presented next, these observations are applied to the case of mathematics learning in Grades 10 through 12. The results of the simulation show that even when there are few links between years, and random patterns of missingness of students, the LG-IRM is able to recover the model parameters well.

Example from LSAY Mathematics Vertical Scale

The results from the simulation provided a solid basis for the application of the LG-IRM to real data. The type of data typically encountered in studies of academic achievement over time include measurements that must be vertically scaled in order to be comparable to eachother. Often this vertical scaling is accomplished through items that are administered in more than one wave of testing. The data selected for this example includes this feature and as such, provides a good example data set for application of the LG-IRM.

Method

Participants.

The example data were selected to resemble the well-behaved vertical scale results from the example. Data were taken from three waves of administration of the mathematics test portion of the Longitudinal Survey of American Youth (LSAY) (Miller, Hoffer, Suchner, Brown, & Nelson, 1992). The students in the first cohort began the LSAY in Grade 10. These students were selected to provide a nationally representative sample of students in 1987 (Miller et al., 1992). As with most longitudinal data collection, not all students were assessed at every time point. In educational data, researchers worry most about participation patterns that are related to the construct being assessed. This can create bias in the sample. One example would be if many students who perform poorly in math eventually drop out of school. Figure 2 shows the participation patterns by grade level. In Grade 10, 2,720 students were assessed. In Grade 11, 551 of those students dropped out from the sample. Not all of these students officially dropped out; 102 of them rejoined the sample in Grade 12. Therefore, only 449 of the original 2,720 students assessed dropped out of the sample and never returned. List-wise deletion would result in a sample of 1,169 students who participated in all three waves of math assessment. Although studies show that missingness patterns should be considered (Diggle, 2002; Newman, 2003), the model will be estimated first using the complete data for simplicity. Students who did not participate in one or more waves were removed from the sample to reduce the amount of missing data. The example already included a large amount of missing data due to the linking across forms and years. The students in the sample were given their first assessment in Grade 10, second in Grade 11, and third in Grade 12. Further research on the LG-IRM could involve patterns of missingness and drop-out.

Forms.

Responses from a bank of 137 mathematics items were selected for the mathematics example. From the bank of 137 items, 77 items were used for Grade 10, 114 items were used for Grade 11, and 106 items were used for Grade 12. This design creates a pattern where some items are present in one administration and then later absent in the next or in combinations thereof. It

is important to note whether the item is present or not in a given wave of administration. These patterns of missingness can be thought of as missing by design and should be reflected in the design matrix. If the design matrix does not correctly model these patterns of missingness, ConQuest will try to estimate more parameters than intended. In most cases, this will result in some parameters being unidentified due to a lack of data. The design matrix for this example includes a column for each estimated item parameter. That column contains possible levels of response (0 or 1) for each wave of administration in which the item is present. In addition, the design matrix constrains the item difficulty to be the same across administrations.

Adding to the complexity of the model specification demands for this particular data, the LSAY designers included three test forms in each year after Grade 10. All students took the same set of items in Grade 10 (Miller et. al., 1992). The later forms were designed to provide differing levels of overall test difficulty and avoid ceiling effects in the assessment program. In the selected sample, students were assigned to either the Medium or Tough form in Grade 11 based on their performance in Grade 10. Again, in Grade 12, students were given a new Easy, Medium, or Tough test form. The form administration by Grade is shown in Table 5. It shows that the majority of students (n=731) took the Grade 10 form, and then took the Tough form in Grades 11 and 12. In order to have been assigned to the Tough test form in both years, these students must have performed well in both Grades 10 and 11. In fact, the LSAY designers noted that an improvement on their design would be to include more difficult items; by Grade 12, students were performing so well that there were significant ceiling effects (LSAY) (Miller et. al., 1992).

Figure 2: Participation Patterns by Grade

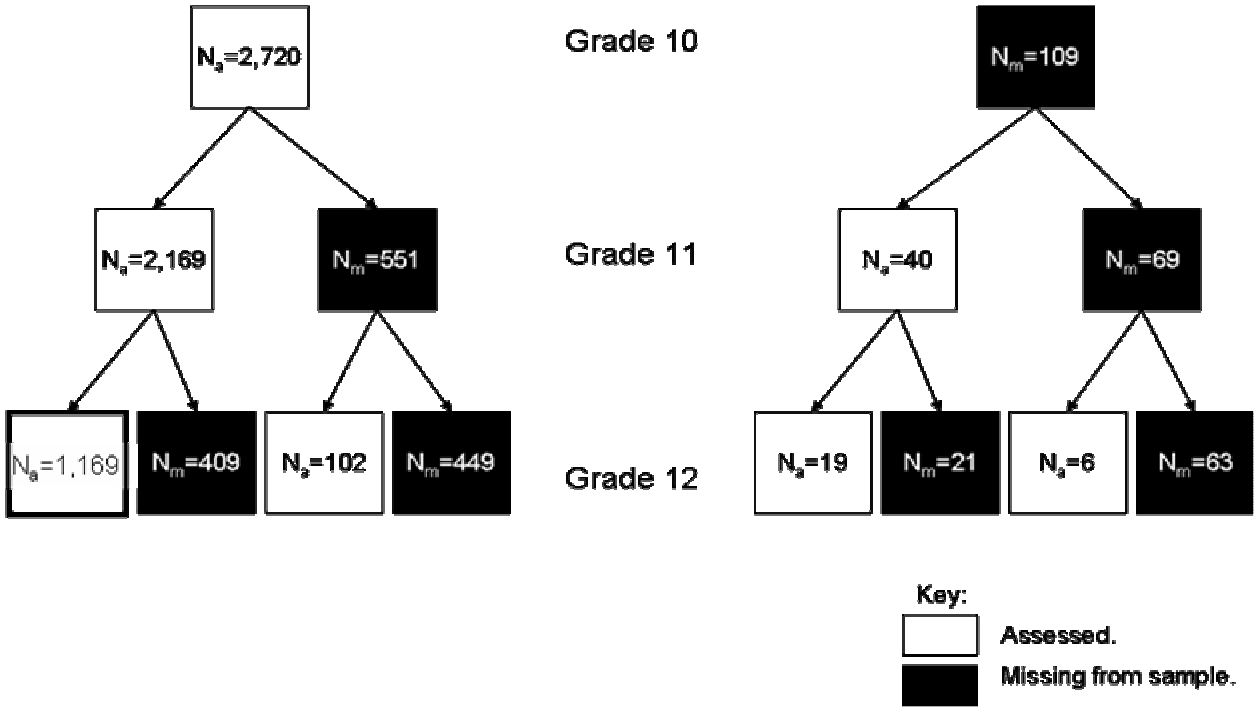


Table 5: Form Administration by Grade

11^{th} Grade Form	12^{th} Grade Form		
Easy	Easy	Medium	Tough
Medium	29	100	130
Tough	11	168	731

Note: N=1,169. Students were placed in form groups based on performance in previous years. All students took the same form in Grade 10.

The data for this example were selected to ensure the largest possible sample size as well as to ensure a sufficient number of common and unique items per year. To ensure a large sample size, the selected students could not be limited to those who took any one pattern of forms. As shown in Table 5, the largest group would have provided a sample size of only 731. Thus, students who took multiple patterns of forms were considered. Given that students in the sample had taken different patterns of forms, which in turn meant that these students took different sets of items, the sample was selected to ensure that some items had been taken by many students. With at least some common items present, the vertical scale was created without estimation problems.

Since the Medium and Tough forms overlapped significantly, it was easy to find items that had been taken by students who took either of these forms in Grade 11 and Grade 12. However, the items on the Easy form in Grade 11, had little overlap with the items on any other form. Therefore, students who took the Easy form in Grade 11 were removed from the sample. This resulted in the final sample size of 1,169 students. In addition, a total of 77 items taken from the Grade 10 form, the Grade 11 Medium and Tough forms, and the Grade 12 Medium and Tough forms could be used for the analysis. These 77 items cover general mathematics, basic mathematics, algebra, quantitative literacy, and geometry. The number and type of item by year is shown in Table 6. In the original LSAY administration, an overall mathematics scale score was produced using all items. In addition, subscale items were used to produce subscale scores in basic mathematics, algebra, quantitative literacy, and geometry. Since the focus of this analysis is to explore growth in mathematics, which was conceptualized as a one-dimensional ability, the aim here is not to report subscale scores. Therefore, items from general mathematics, basic mathematics, algebra, quantitative literacy, and geometry were combined and used to estimate mathematics ability. This method is consistent with the way that the overall mathematics score was calculated in the original LSAY administration (Miller et. al., 1992).

Table 6: Number of Items in Each Area by Year

	10th Grade (Year 0)	11th Grade (Year 1)	12th Grade (Year 2)
Basic Math	20	13	26
Quant. Lit.	14	9	14
Algebra	8	3	8
Geometry	17	12	17
General Math	10	4	12
Total	69	41	77

The items used for each subscale in the original LSAY administration cover different topics. Table 6 shows the number of items in each area for each year. The bottom row of the table shows the total number of items in each area that were disbursed across the three forms. A total of twelve general mathematics items cover basic manipulation and order of operations. A total of 26 basic mathematics items cover basic manipulation and functions. The eight algebra items cover unit conversions, solving equations for one unknown, and inequalities. The 24 basic mathematics items require students to understand fractions, the manipulation of currency values, rounding, and negative values. The seventeen geometry items cover the geometry of triangles, circles, arcs, and rectangles. In many of the items, students are asked to find the angle or length of a side given, other angles, or side lengths. The quantitative literacy area covers principles of probability. These fourteen items include questions about combinations, permutations, and probability. Items in all subscales are presented using different mathematic representations. These items are presented using equations, graphical tools such as number lines, and simple story-problem scenarios. These features may influence the difficulty of the items or their bias. Since these secondary data do not include access to information on the features and formats that were intentionally designed into the items, no formal analysis of items properties as predictors of item difficulty will be explored. However, in the next chapter, features of these items, as potential sources of bias, will be explored.

Example Results

ConQuest estimated the LG-IRM with three waves of data. A design matrix was created to map the items to the baseline and growth dimension (Figure 3). In Figure 3, design blocks for each year are stacked on top of each other to match up the item parameters in each year. The number of rows in each design block is determined by the number of items presented in that form. For instance, in Year 1, only 41 items were used, so this design block is shorter than the others. The columns specify the 76 item parameters to be estimated. This design matrix was specifically formatted to the 77 items selected for this analysis. In order to make estimation possible, a constraint was placed on the 77^{th} item. Therefore, difficulty for that item was not estimated. In addition, estimates of the population parameters were produced. These included the means and variances of the baseline ability and growth parameter, as well as the correlation between the two. Once the item parameters and population parameter estimates were produced, the WLEs (Warm, 1989) of the baseline ability and growth parameter for each individual were obtained.

Item parameters.

The full set of item parameter estimates is shown in Table 7. The Unweighted fit values shown include the Mean Square (MNSQ), its 95% confidence interval, and the t-statistic. The Weighted fit values shown include the MNSQ, its 95% confidence interval, and the t-statistic. The unweighted fit is sometimes referred to as *infit* and the weighted fit is sometimes referred to as *outfit*. The unweighted MNSQ is calculated by squaring and averaging the standardized residuals. The weighted MNSQ is calculated by squaring and averaging the standardized residuals with weights (Bond & Fox, 2001; Wright & Masters, 1982). Thus the MNSQ are the same as a Chi-squared statistic divided by its degrees of freedom. The t-statistic reports the significance of the MNSQ, calculated as a one would calculate a z-statistic. However in the Rasch modeling literature, these are referred to as t-statistics. The combined fit values provide

evidence of item fit and dimensionality. The dimensionality evidence is provided in terms of violations of the unidimensionality assumption. For instance if the group of quantitative literacy items all showed poor fit, this would be evidence that those items are part of a different dimension. Since these LSAY items were originally NAEP items, it is not surprising that the unweighted and weighted fit demonstrate good fit for all of the items. The unweighted and weighted fit values are all within the acceptable range of 0.75 to 1.33 (Bond & Fox, 2001). This gives evidence that the item difficulties fit well.

Figure 3: Design matrix for analysis of three waves of data using the LG-IRM

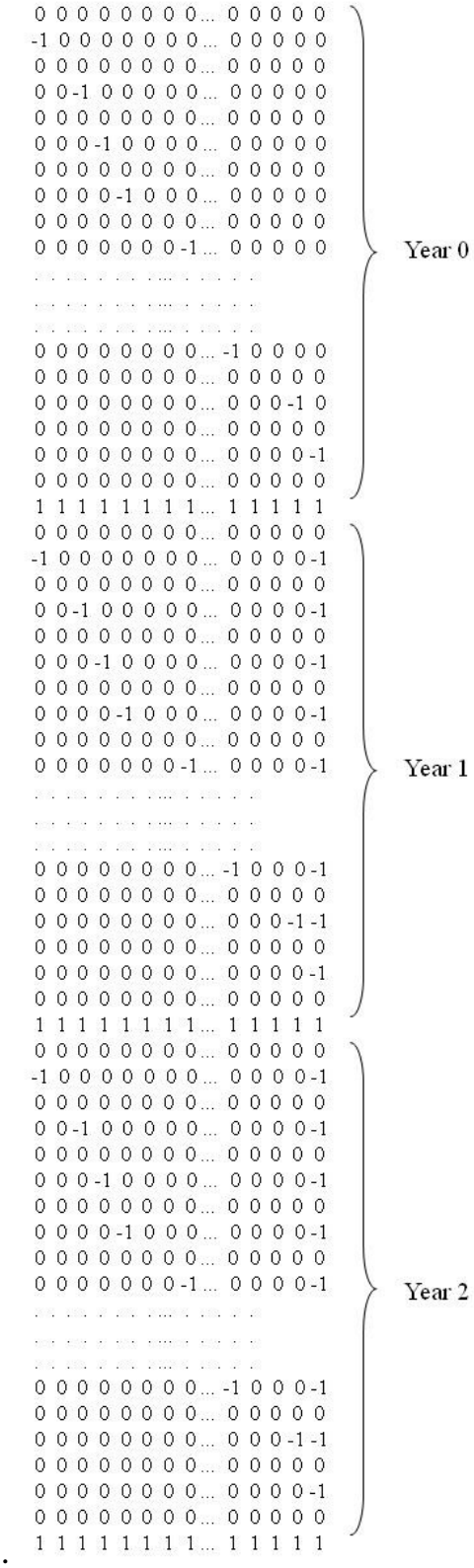


Table 7: Mathematics Example Item Parameter Estimates

			Unweighted Fit			Weighted Fit		
Parameter	Difficulty	SE	MNSQ	95% C.I.	T	MNSQ	95% C.I.	T
1	0.6156	0.0348	1.10	(0.92, 1.08)	2.3	1.21	(0.91, 1.09)	4.3
2	1.5545	0.0444	0.93	(0.92, 1.08)	-1.7	1.04	(0.91, 1.09)	0.9
3	0.3696	0.0335	1.24	(0.92, 1.08)	5.4	1.38	(0.91, 1.09)	7.6
4	0.1547	0.0326	1.05	(0.92, 1.08)	1.1	1.12	(0.91, 1.09)	2.6
5	-0.3322	0.0313	1.03	(0.92, 1.08)	0.7	1.12	(0.92, 1.08)	2.8
6	1.4895	0.0463	0.89	(0.92, 1.08)	-2.8	1.04	(0.90, 1.10)	0.9
7	1.0497	0.0497	0.94	(0.92, 1.08)	-1.5	1.09	(0.91, 1.09)	1.9
8	0.2525	0.0380	1.16	(0.92, 1.08)	3.7	1.16	(0.91, 1.09)	3.4
9	-0.2353	0.0313	1.17	(0.92, 1.08)	3.8	1.17	(0.92, 1.08)	3.8
10	1.3050	0.0423	0.90	(0.92, 1.08)	-2.4	1.07	(0.91, 1.09)	1.5
11	-0.0771	0.0317	1.14	(0.92, 1.08)	3.3	1.21	(0.92, 1.08)	4.5
12	1.1536	0.0417	0.90	(0.92, 1.08)	-2.4	1.00	(0.91, 1.09)	0
13	0.1892	0.0327	1.05	(0.92, 1.08)	1.2	1.19	(0.91, 1.09)	4.1
14	0.2170	0.0328	1.21	(0.92, 1.08)	4.7	1.24	(0.91, 1.09)	5
15	-1.5637	0.0300	1.31	(0.92, 1.08)	6.8	1.30	(0.92, 1.08)	7
16	0.7072	0.0374	0.84	(0.92, 1.08)	-4.1	0.94	(0.91, 1.09)	-1.3
17	-0.7765	0.0351	1.12	(0.92, 1.08)	2.9	1.14	(0.92, 1.08)	3.4
18	-0.5506	0.0332	1.12	(0.92, 1.08)	2.8	1.16	(0.92, 1.08)	3.7
19	-0.0274	0.0320	1.33	(0.92, 1.08)	7.2	1.35	(0.91, 1.09)	7.4
20	-1.9181	0.0319	1.25	(0.92, 1.08)	5.5	1.25	(0.92, 1.08)	6
21	0.9492	0.0363	0.98	(0.92, 1.08)	-0.6	1.16	(0.91, 1.09)	3.2
22	-1.1158	0.0300	1.12	(0.92, 1.08)	2.9	1.18	(0.92, 1.08)	4.2
23	0.5245	0.0342	1.00	(0.92, 1.08)	0.1	1.09	(0.91, 1.09)	2
24	-1.3045	0.0313	1.37	(0.92, 1.08)	8	1.38	(0.92, 1.08)	8.7
25	-0.9445	0.0299	1.27	(0.92, 1.08)	6	1.26	(0.92, 1.08)	5.9
26	-0.2381	0.0339	0.96	(0.92, 1.08)	-1	1.00	(0.91, 1.09)	0
27	-0.0140	0.0320	0.96	(0.92, 1.08)	-1	1.06	(0.91, 1.09)	1.3
28	-0.2889	0.0311	1.08	(0.92, 1.08)	2	1.16	(0.92, 1.08)	3.5
29	-0.3201	0.0335	1.01	(0.92, 1.08)	0.4	1.11	(0.91, 1.09)	2.4
30	0.1483	0.0349	1.06	(0.92, 1.08)	1.4	1.20	(0.91, 1.09)	4.1
31	-2.1454	0.0323	1.31	(0.92, 1.08)	6.9	1.31	(0.92, 1.08)	7.2
32	-1.5432	0.0315	1.12	(0.92, 1.08)	2.7	1.12	(0.92, 1.08)	3
33	1.0479	0.0413	0.86	(0.92, 1.08)	-3.6	1.00	(0.91, 1.09)	0.1
34	0.5027	0.0362	0.89	(0.92, 1.08)	-2.7	1.05	(0.91, 1.09)	1.2
35	0.4364	0.0359	0.96	(0.92, 1.08)	-0.9	1.10	(0.91, 1.09)	2.1
36	0.9443	0.0407	1.02	(0.92, 1.08)	0.5	1.09	(0.91, 1.09)	2
37	0.4873	0.0362	1.40	(0.92, 1.08)	8.5	1.27	(0.91, 1.09)	5.3
38	1.5231	0.0432	1.04	(0.92, 1.08)	0.9	1.14	(0.91, 1.09)	2.8
39	-0.8265	0.0318	1.22	(0.92, 1.08)	5	1.21	(0.92, 1.08)	4.9

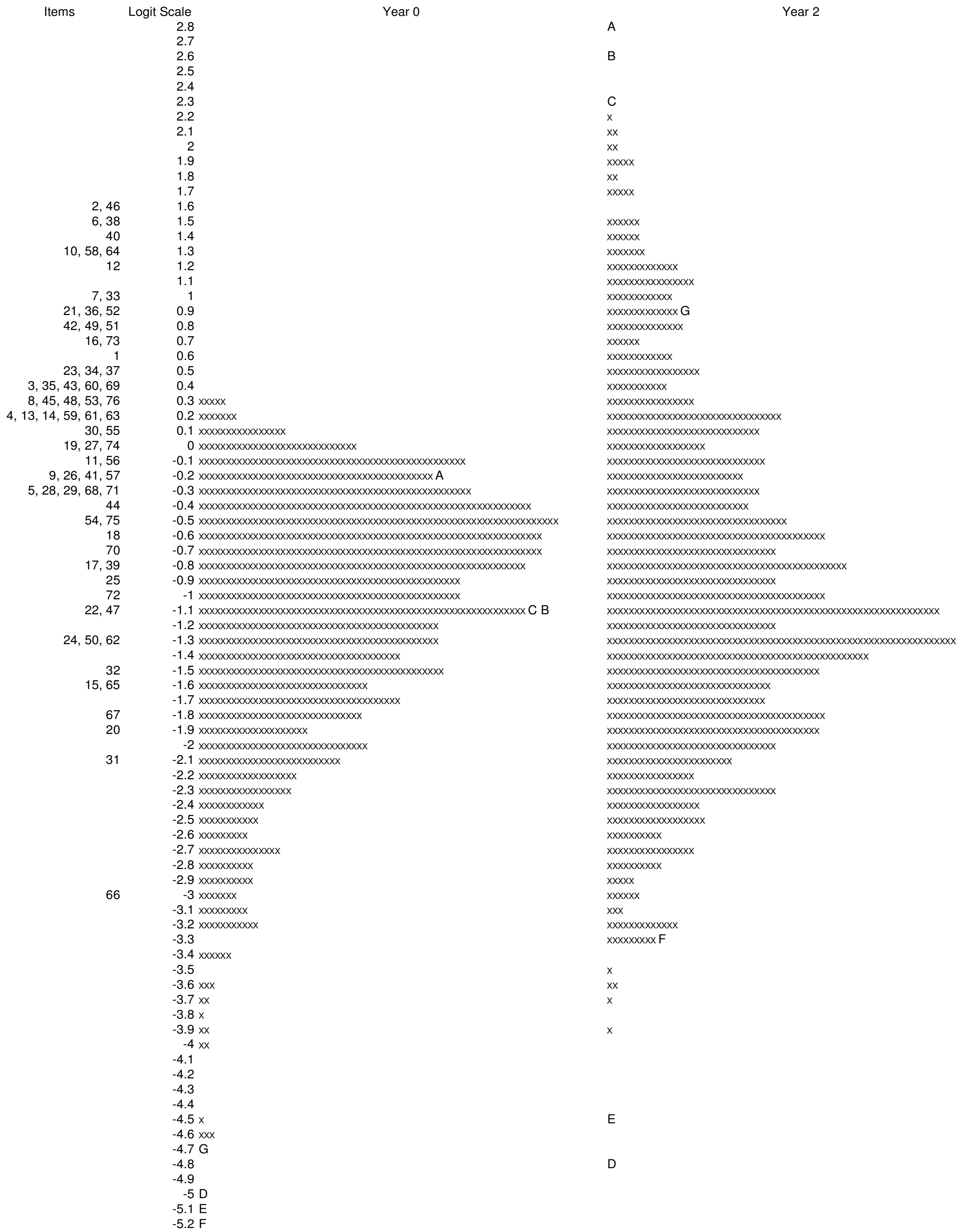
Table 7 (cont.): Mathematics Example Item Parameter Estimates

			Unweighted Fit			Weighted Fit		
Parameter	Difficulty	SE	MNSQ	95% C.I.	T	MNSQ	95% C.I.	T
41	-0.1847	0.0314	1.14	(0.92, 1.08)	3.3	1.23	(0.92, 1.08)	5.1
42	0.7727	0.0374	1.00	(0.92, 1.08)	0.1	1.17	(0.91, 1.09)	3.3
43	0.3928	0.0336	0.93	(0.92, 1.08)	-1.8	1.01	(0.91, 1.09)	0.3
44	-0.4188	0.0326	1.13	(0.92, 1.08)	3.1	1.17	(0.92, 1.08)	3.9
45	0.275	0.0379	0.94	(0.92, 1.08)	-1.4	1.08	(0.91, 1.09)	1.7
46	1.6086	0.0497	0.94	(0.92, 1.08)	-1.6	1.08	(0.91, 1.09)	1.8
47	-1.1447	0.0315	1.18	(0.92, 1.08)	4	1.23	(0.92, 1.08)	5.5
48	0.2637	0.033	1.13	(0.92, 1.08)	3	1.22	(0.91, 1.09)	4.6
49	0.8115	0.0403	0.99	(0.92, 1.08)	-0.3	1.11	(0.91, 1.09)	2.4
50	-1.2759	0.0313	1.18	(0.92, 1.08)	4.1	1.19	(0.92, 1.08)	4.6
51	0.8387	0.0402	1.03	(0.92, 1.08)	0.6	1.11	(0.91, 1.09)	2.3
52	0.8568	0.0359	1.10	(0.92, 1.08)	2.4	1.24	(0.91, 1.09)	4.8
53	0.2949	0.0331	1.11	(0.92, 1.08)	2.5	1.28	(0.91, 1.09)	5.7
54	-0.5467	0.0357	1.18	(0.92, 1.08)	4.2	1.2	(0.92, 1.08)	4.6
55	0.062	0.0374	1.2	(0.92, 1.08)	4.5	1.15	(0.91, 1.09)	3.3
56	-0.051	0.0368	0.97	(0.92, 1.08)	-0.7	1.05	(0.92, 1.08)	1.2
57	-0.2346	0.0432	1.02	(0.92, 1.08)	0.4	1.06	(0.91, 1.09)	1.4
58	1.3223	0.0497	0.94	(0.92, 1.08)	-1.5	1.09	(0.91, 1.09)	2
59	0.1527	0.0375	1.02	(0.92, 1.08)	0.4	1.11	(0.91, 1.09)	2.5
60	0.3767	0.0384	0.89	(0.92, 1.08)	-2.8	0.98	(0.91, 1.09)	-0.4
61	0.2378	0.038	0.94	(0.92, 1.08)	-1.5	1.02	(0.91, 1.09)	0.5
62	-1.2853	0.0348	1.17	(0.92, 1.08)	4	1.16	(0.92, 1.08)	3.9
63	0.2123	0.0377	0.98	(0.92, 1.08)	-0.4	1.07	(0.91, 1.09)	1.6
64	1.3052	0.0497	0.94	(0.92, 1.08)	-1.6	1.08	(0.91, 1.09)	1.8
65	-1.5963	0.035	1.19	(0.92, 1.08)	4.4	1.2	(0.92, 1.08)	4.7
66	-2.9998	0.0433	1.18	(0.92, 1.08)	4.1	1.21	(0.92, 1.08)	4.9
67	-1.7697	0.0354	1.13	(0.92, 1.08)	3.1	1.14	(0.92, 1.08)	3.4
68	-0.3313	0.036	0.94	(0.92, 1.08)	-1.5	1.02	(0.92, 1.08)	0.5
69	0.4146	0.0384	1.01	(0.92, 1.08)	0.4	1.07	(0.91, 1.09)	1.5
70	-0.6687	0.0354	1.17	(0.92, 1.08)	3.9	1.23	(0.92, 1.08)	5.2
71	-0.3358	0.0359	1.16	(0.92, 1.08)	3.8	1.22	(0.92, 1.08)	4.9
72	-1.033	0.0351	1.15	(0.92, 1.08)	3.6	1.15	(0.92, 1.08)	3.7
73	0.6501	0.0395	1.01	(0.92, 1.08)	0.3	1.03	(0.91, 1.09)	0.7
74	-0.0094	0.0369	0.89	(0.92, 1.08)	-2.8	0.99	(0.92, 1.08)	-0.2
75	-0.4751	0.0357	1.19	(0.92, 1.08)	4.2	1.16	(0.92, 1.08)	3.8

However, this is not the case. So it is reasonable to assume unidimensionality for the example data. Domains and item formats can be used as an informal way to explore why an item may have been more or less difficult. The easiest item, Item 66, is a General Math item. It asks students to surfaces of a 3D shape when that 3D shape is sliced. The most difficult item, Item 2, is a basic mathematics item. It is presented as a word problem that also requires the student to reduce a fraction. It is possible that the combination of the word problem format and fraction reduction makes the item difficult. Item 46 is another one of the most difficult items. It is also a basic mathematics problem formulated as a word problem. The item difficulties are also represented on the Annotated Wright Map shown in Figure 4.

The Annotated Wright Map combines representation of the item difficulties, item subscales, and personal estimates on the logit scale. As such, it allows for comparison between item difficulties and personal abilities. The item and person correspondence from the LG-IRM can be interpreted in the same way as other models of the Rasch family. The interpretation is that if a personal ability estimate has the same value as a particular item difficulty, that person has a 50% chance of answering that item correctly. If the personal ability estimate is above the item difficulty, then that person has more than a 50% chance of answering the item correctly. The opposite applies when the personal ability estimate is below the item difficulty. Although some of the LG-IRM analysis in ConQuest is very similar to other Rasch family analysis, it also requires additional user manipulation. One example is the design matrix, as discussed earlier. In addition, the Wright map must be produced manually. This is because using the design matrix prevents ConQuest from producing the Wright Map. Future work in this area might include developing tools to produce the Wright Maps automatically.

Figure 4: Annotated Wright Map showing difficulty of items, baseline ability, and ability at time 2 using the ability estimates from the LG-IRM with example cases highlighted with letters A-G.



Growth in mathematics ability.

The results show that on average students achieved a small but positive amount of growth in mathematics from Grade 10 to 12, as measured by the LSAY items. The mean baseline ability

(-1.128 logits) and mean growth (0.058 logits per year) are shown in Table 8. The standard errors of each estimate are also given.

Table 8: Pearson Parameter Estimates

Dimension	Mean	SE	Variance
Baseline	-1.128	0.023	0.603
Growth	0.058	0.013	0.184

Note. Correlation between baseline and growth was small but positive (p=.028). Items from the Grade 10 form were used to define the baseline dimension. Items from all three waves were used to define the linear growth parameter.

Together, the estimates and standard errors suggest that the average student would achieve growth of around 0.058 logit per year in mathematics ability. In addition, it was found that the baseline ability and growth parameter are positively correlated (0.028). The message that students achieve small amounts of positive growth is also seen visually in the average shift upward of the distributions in the Wright Map (Figure 4).

The manually produced Wright Map in Figure 4 was also used to highlight individual growth trajectories of interesting students. The student locations at the baseline and by Year 2, or Grade 12, are labeled with letters A-G. Student A ends with the highest ability at time 2. Students B and C also have abilities far above the rest of the student distribution at Grade 12. However, these students had different baseline abilities and growth rates. For instance, although Student A has a much higher baseline ability than Student B, Students A and B end with abilities that are much closer together. This is because Student B has a fast growth rate. This is also the reason why Student B ends up with a much higher ability than Student C even though they start at the same baseline ability.

Student B, then, represents a good example of student who starts at the high end of the distribution and continues to learn quickly. This student's course-taking patterns reveal that Student B took classes through to level two of algebra by the end of Grade 10. TThis student then took mathematics courses in a less traditional track, including vocational mathematics, business mathematics, and computers. This student may have achieved a steep rate of growth through being placed in this track. However, the student with the steepest learning trajectory is labeled "G." This student took level two of algebra in Grade 10, analytic geometry, as well as trigonometry and pre-calculus in Grade 11 and calculus in Grade 12. Assumedly, this student did well in all of these courses. Thus, the Wright Map can be used to identify individual student trajectories within the distributions of student abilities across time. If the data include sufficient numbers of students clustered around each teacher, student learning trajectories could also be linked to teachers as well. Such links might provide information on teacher quality.

Group differences.

Issues surrounding the achievement gap can also be explored using the estimates from the LG-IRM. One common question is whether certain groups of students learn at a faster rate. If only the IRT scores are available, a first step might be to conduct a t-test, comparing the mean baseline ability for men and women. The null hypothesis is that boys and girls have equal mathematics abilities at baseline. The WLE estimates of each student's baseline ability were used for this test, the results of which are shown in Table 9. The t-test shows that the null hypothesis should be rejected ($t(1167)=8.113$, $p < .01$). This result may mean that men and women do differ in baseline ability. The same t-test was performed again using the WLE

estimates of the growth parameter. The null hypothesis in this case is that boys and girls have the same growth rate. The results of this test (Table 9) show that the null hypothesis should not be rejected in this case ($t(1167)=-1.434$, $p = 0.152$).

The combined results of these two t-tests suggest that boys and girls start at different levels of mathematics ability in Grade 10, but that one group does not learn faster than another. Thus, if the boys start at a lower level than the girls, they will not be able to catch up by Grade 12 by having a faster growth rate. Another way to explore the question of group differences might be to include gender as a covariate in the latent regressions for baseline ability and growth. This may be a topic for further exploration.

Table 9: t-test by Gender

Group	Female	Male	Difference
N	581	588	
$Mean_1$	-0.958	-1.368	0.41
SE_1	0.029	0.041	0.051
SD_1	0.702	0.999	
95% CI_1	(-1.015, -0.900)	(-1.449, -1.287)	(0.311, 0.509)
t_1			8.113**
$Mean_2$	0.082	0.12	-0.386
SE_2	0.017	0.211	0.027
SD_2	0.401	0.511	
95% CI_2	(0.049, 0.115)	(0.079, 0.161)	(-0.091, 0.014)
t_2			-1.434

Note: The t-test using the baseline estimate was significant, t(1167), p<.01. The t-test using the linear growth parameter was not significant, t(1167), p=0.152.

Although it is interesting to explore group differences based on student demographics, it is practical to look at features of the educational experience as well. For instance, if course taking patterns are found to be related to student growth, additional attention could be given to planning and studying course pathways to promote student learning. The LSAY combines academic testing with additional information taken from interviews with students and parents. The spring student interviews ask students to speak about their courses and teachers. From these interviews, the LSAY administrators were able to obtain information on student course-taking patterns in mathematics. By combining these reports with the results from the LSAY, it was possible to explore whether course-taking patterns were related to the student learning trajectories. The courses ranged from geometry and algebra to calculus, and also included courses such as business mathematic and statistics. Given that courses taken in each semester were also reported, unique course-taking patterns could be examined for their relationship to the estimated learning trajectories in mathematics. Since there were over 300 course-taking patterns, only some were selected for discussion here, but a more systematic study of course-taking patterns and student could be a topic for further research. Figure 5 illustrates how individual students can achieve at the same level by Grade 12 even after starting at very different baseline abilities. For example, one student who took algebra II in both the fall and spring terms of Grade 10, and no further mathematics courses except geometry in the spring of Grade 11 (denoted with a gray +), was able to catch up with a student who took geometry in both the fall and spring terms of Grade 10,

algebra II in the fall of Grade 11, geometry in the spring of Grade 11, and statistics/probability in the fall of Grade 12 (denoted with gray squares) even though they had started at very different abilities levels in the fall of Grade 10. Other students had very similar growth trajectories even though they took different courses (students denoted by the dashed line and black "x"). The student taking consumer mathematics and vocational courses (denoted with the gray circles) maintained the highest ability over time, reflected by what was measured by the items on the test. The differences between the individual growth trajectories shown in Figure 5 illustrate how accountability designs based on a set target of achievement in terms of a score on some scale, could fail to identify students who have achieved large amounts of growth if there progress is only reported in terms of whether or not a target ability level was met.

If broader conclusions are to be drawn for the course-taking patterns in the LSAY, it might be more helpful to use a summary of the courses taken for groups of students rather than the individual course progressions. The highest course taken through the end of each grade was also reported (Table 10) and summarizes the courses taken. By the end of Grade 10, most students had taken mathematics courses through to geometry or geometry honors (42.6%, 2.74%) or through algebra II or algebra II honors (23.61%, 4.19%). An additional, smaller group had taken courses through to algebra I or algebra I honors (16.08%, 0.68%). By the end of Grade 11, most students had taken courses through to algebra II or algebra II honors (34.39%, 1.46%) or through to geometry or geometry honors (19.43%, 1.38%). By the end of Grade 11, there were also groups of students who had stopped taking mathematics courses (11.18%). Finally, by the end of Grade 12, an even larger group of students was no longer taking mathematics courses (37.48%). Students were also taking calculus (15.01%), and analytic geometry and pre-calculus (14.07%). These summary variables show the differences in baseline ability for groups defined by the highest math courses taken in each grade. Figure 6 shows different baseline abilities, or the ability calculated for the fall term of Grade 10 for each group. The slopes or growth rates are fairly similar. The representation in Figure 6 is one way to describe the differences in the estimated baseline ability and student growth using the highest mathematics courses taken in each year. A similar display could be constructed for students of different schools, teachers, and classrooms, as well to further describe growth in mathematics growth in terms of related inputs like courses taken and teacher quality.

Figure 5: Individual trajectories by course-taking pattern reported in the five terms including the fall of Grade 10, the spring of Grade 10, the fall of Grade 11, the spring of Grade 11, and the fall of Grade 12.

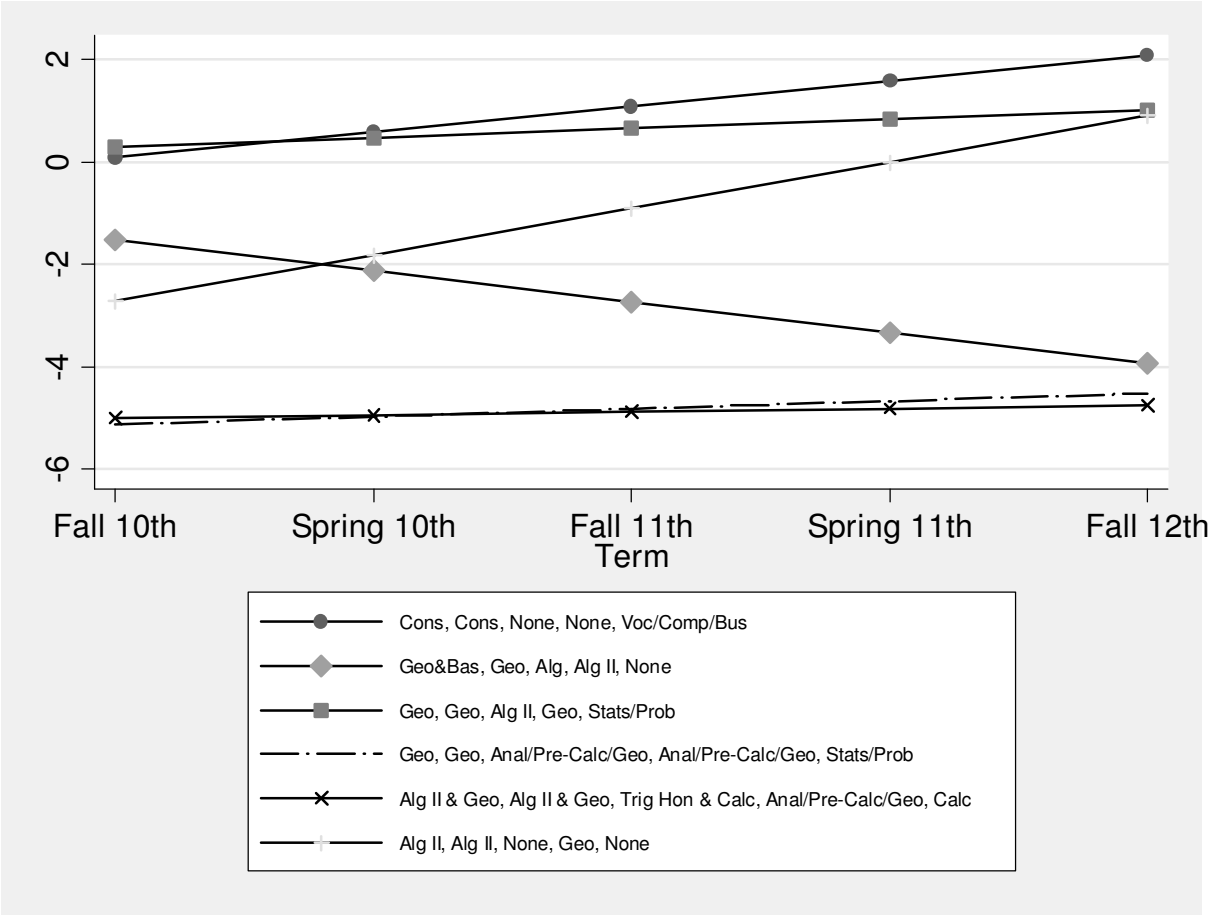


Table 10: Course-taking Patterns

	Grade 10	Grade 11	Grade 12
No Course	17 (1.45%)	130 (11.18%)	437 (37.48%)
Basic Math	30 (2.57%)	5 (0.43%)	5 (0.43%)
Vocational Math/Computers/Business	6 (0.51%)	28 (2.41%)	32 (2.74%)
Consumer Math	28 (2.40%)	34 (2.92%)	18 (1.54%)
Geometry	498 (42.6%)	226 (19.43%)	41 (3.52%)
Honors Geometry	32 (2.74%)	16 (1.38%)	
Pre-Algebra	22 (1.88%)	5 (0.43%)	5 (0.43%)
Algebra 1	188 (16.08%)	42 (3.61%)	23 (1.97%)
Algebra 1 Honors	8 (0.68%)		
Algebra 2	276 (23.61%)	400 (34.39%)	138 (11.84%)
Algebra 2 Honors	49 (4.19%)	17 (1.46%)	4 (0.34%)
Trigonometry	7 (0.60%)	90 (7.74%)	102 (8.75%)
Trig Honors	1 (0.09%)	36 (3.10%)	2 (0.17%)
Analytic Geometry/Pre-Calc	7 (0.60%)	121 (10.4%)	164 (14.07%)
Calculus		8 (0.69%)	175 (15.01%)
Statistics/Probability		2 (0.17%)	14 (1.20%)
Not in School		3 (0.26%)	6 (0.51%)

Note: In Grade 10, all students (N_{10}=1,169) reported course taking patterns. In Grades 11 and 12, the number of course taking patterns (N_{11}=1,158, N_{12}=1,166) were slightly smaller.

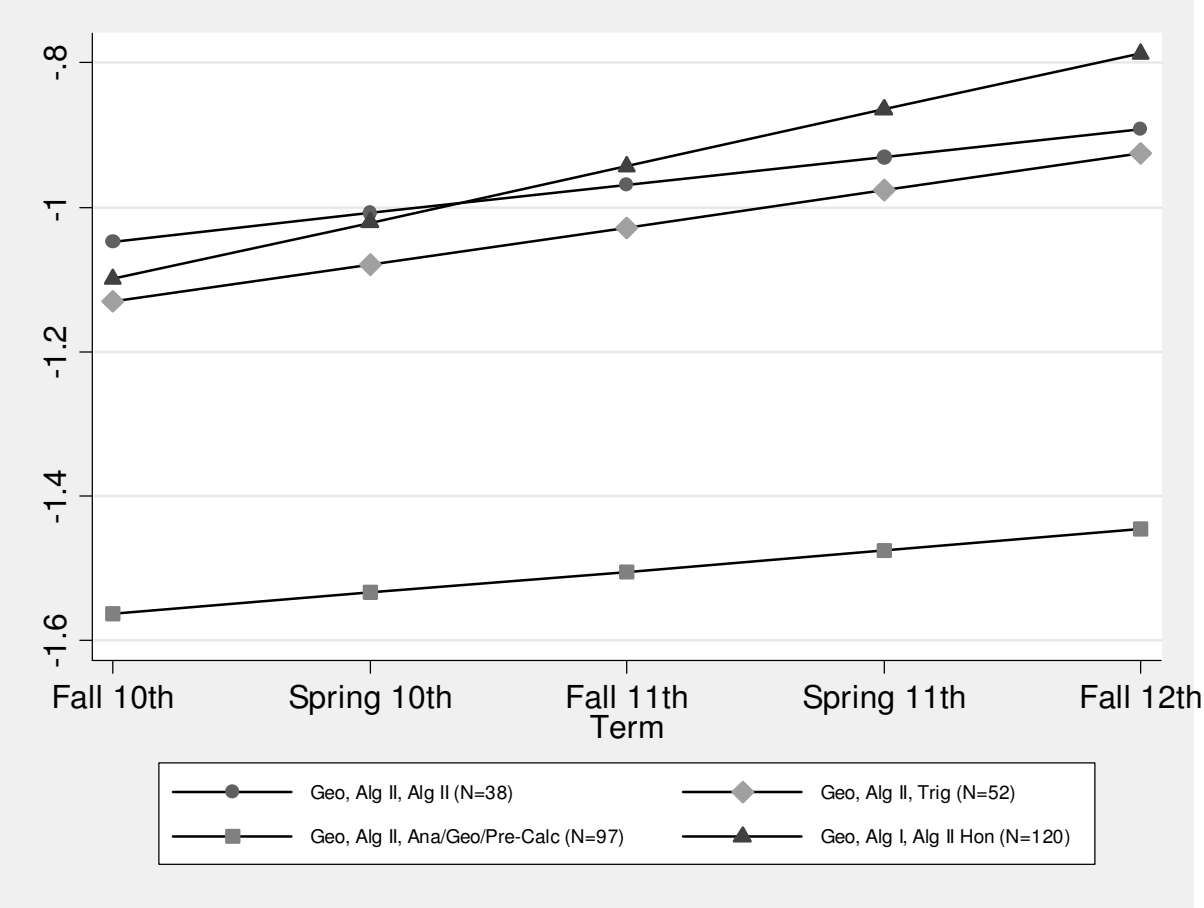
Discussion

This paper presented a simulation example of a vertically scaled assessment in which the LG-IRM was employed. In the simulation example, the LG-IRM parameters were well-recovered. The simulation design and estimation was described in detail to illustrate important components of the LG-IRM, including the way that the vertical scale is created. The results of the simulation evidence that the LG-IRM can be employed when complex mappings between the item parameters and latent dimensions are required.

Next the data was estimated using real data from a longitudinal study which included a mathematics test in each year. Examples of representations such as the Wright Map and example growth trajectory plots were illustrated. The Wright Map was used to illustrate the shift in distribution of learning from baseline to time 2. It was also used to draw connections between the amount of learning and the items that were used to assess it. Example growth trajectories were used to provide visual representations of the differences in slope and baseline ability of individuals and groups of students. The representations used to explain the example results were created to show how the results of the LG-IRM might provide meaningful information on how students are learning.

Finally, the example results were analyzed for group differences. The results from the mathematics example show differences in baseline ability and linear growth rate by group characteristics. Patterns were found to be related to gender. Multiple societal mechanisms may be responsible for these differences related to gender. A particular item, for instance may favor some students over others based on race or gender. If this is the case, psychometricians can help

Figure 6: Mean growth trajectories by highest mathematics course taken in Grade 10, Grade 11, and Grade 12.



by flagging or removing these items. This small step can help to ensure that the instruments used to assess students are fair. Patterns in growth were also found using student reports on the mathematics courses taken. The patterns in learning by course taking could be related to the material taught in the course. For instance, a course such as geometry taken in a certain year may be easier for those students than the geometry items. Or they may become more difficult as memory effects take effect. Thus, these group differences may also be related to curriculum-based item shifts. These could be conceptualized as curriculum DIF parameters. The differences could also be related to ability and motivation, since this may effect course selection and assignment. Analyses of group differences are often conducted using DIF. The incorporation of DIF into the LG-IRM could be a topic of future exploration.

Chapter 2: Changes in Low Self-Esteem Over Time: Evidence from the National Survey of Black Americans

"There is probably no personality trait more significant in the context of total psychological functioning than self-esteem" (Kawash & Scherf, 1975, p. 715).

Self-esteem is an important correlate of many factors of personal wellbeing. As such, it has been measured and studied in a variety of research domains. Often, the self-esteem of Blacks is compared to that of Whites (e.g. Simmons, 1978). Despite racial discrimination, Blacks tend to report higher level of self-esteem than Whites (Gray-Little & Hafdahl, 2000; Tweng & Crocker, 2002). Several studies explored the possible reasons behind the reporting of high self-esteem among Blacks (e.g. Hughes & Demo, 1980). The scholarly community has yet to arrive at a definitive answer as to why Blacks report higher levels of self-esteem than whites, but several promising theories have emerged. Several factors could explain the differences between Black and White self-esteem. Some find that ethnic identity is an important correlate of self-esteem (Twenge & Crocker, 2002; Phinney & Chavira, 1995). Others find that Blacks and Whites actually draw on different domains of reference when reporting on the self-esteem scale of respondents (Crocker, Luhtanen, Cooper & Bouvrette, 2003). In addition, some consider self-esteem to be contingent on success in certain domains. These domains range from academic success (Baumister, 2005; Crocker, Karpinski, Quinn, & Chase, 2003; Crocker, Sommers, & Luhtanen, 2002) to athletic success (e.g. Prasad & Thakur, 1977) and social success (Crocker et. al., 2003). Thus, increases in success in any one of these domains could explain increases in self-esteem (Crocker & Park, 2004; Crocker & Wolfe, 2004). These contingency studies also draw on the important point that self-esteem can change over time.

It has been theorized that self-esteem is relatively *stable* over time. However, as noted in *contingency* theory, some experiences can lead to change in Black and White self-esteem. Also, during certain developmental periods, self-esteem is expected to change. For instance, one study found that for students less than 10 years old, Blacks actually report lower levels of self-esteem that Whites. With maturity, the pattern of Blacks reporting higher levels of self-esteem than Whites re-appears (Robins, Trzesniewski, Tracy, Gosling, & Potter, 2002). When examining changes in the self-esteem of both Blacks and Whites over time, a few general patterns have been noted. First, it appears that self-esteem decreases from childhood through adolescence. Then self-esteem slowly increases during adulthood. Finally, self-esteem decreases with old-age. It has also been hypothesized that Black self-esteem has increased during certain periods in U.S. history (Twenge & Crocker, 2002). Scholars have theorized that the election of the first African-American president would provide a boost to Black self-esteem as well, but no longitudinal studies have yet been conducted to explore this theory.

Using the scores reported by the self-esteem measure, the researcher analyzes differences in levels of self-esteem. However, differences and changes in levels of self-esteem can be difficult to interpret if the connections between these amounts are not made explicit in terms of the self-esteem construct. It has been argued that the scope of research on self-esteem must be broadened so that these levels can be understood and interpreted (Zeigler, 2007). Studies of this nature, if modeled adequately, could provide information on how much Black self-esteem changes over the life-course and during certain historical periods, as well as how to interpret this change in terms of the self-esteem construct that was measured.

The Present Study

This study will quantify improvements in low self-esteem over time, as measured by the National Survey of Black Americans (NSBA) (Jackson & Neighbors, 1996). Using available tools of the Item Response Modeling (IRT) Framework, such as the Wright Map, results will demonstrate how meaningful interpretations of the amount of improvement in low-self esteem could be made. Building the growth model will consider other aspects as well, including how to properly model the likert-type response options and how to model the participation patterns of the respondents. This study provides an example of how the Latent Growth Item Response Model can be applied to personality research and to polytomous data.

IRT Methodologies in Personality Research

Personality research, including research on self-esteem, has a long history of using advanced statistical models and methods to construct, validate, and interpret its measures. In order to validate the structure of the constructs being measured, researchers often employ factor analytic methods. Regression techniques and correlational analysis are also often used. To study changes over time, longitudinal studies of self-esteem scores have also been conducted (Rapkin & Hirsch, 1987; Block & Robins, 1993; Zimmerman, Copeland, & Dielman, 1997). A more recent trend has been to employ IRT modeling techniques in personality research. This trend is making its way into the community through scholarly work being published in the measurement journals, and sprinkled in the psychological journals (Meijer, 2002). Some these applications of IRT relate specifically to studies of self-esteem. For instance, the Graded Response Model has been applied to a study of self-esteem using the Rosenberg Scale (1965) (Gray-Little, Willams, & Hancock, 1997). In addition, mixture IRT models have been employed to search for latent classes of examinees (Rost, 1990, 1991; von Davier, 1995). The use of IRT for personality research affords the means for comparison of the response to the construct being measured (Steinberg & Thissen, 1996). It also affords the chance for change to be interpreted in terms of the construct being measured. However, several differences between measurement in personality research and educational research should be noted before an IRT model is applied.

The instrument design.

If the analysis is to be conducted on secondary data, researchers should know the construction of the instrument. The wording and format of the items influences the decision on which IRT model to employ and how to prepare the data for analysis. For instance, in practical personality research, the length of the measure may be abbreviated so that it may be administered in a short amount of time. In this case, in order to achieve high reliability, there may be an oversampling of items in a particular topic. The Rosenberg Scale (1965), for instance, includes both positively and negatively worded questions on the same topic. The target level of assessment, or the level at which the assessment is designed to provide the most information, may also differ in personality research. A test of ability might be designed to measure student abilities across a span of a few grades. Thus, the items would be chosen from material covered in all of the grades, and span a broad range of knowledge. A self-esteem measure may distinguish between respondents with low self-esteem, who might be at risk for negative outcomes. It will have a lot of items targeted at fairly extreme characteristics for the broader population. Differences may also exist in the response options. For instance, large-scale educational tests tend to consist of multiple-choice questions with distracters. Personality instruments tend to include 4-6 Likert-type response options, although some argue that respondents probably cannot

make so many distinctions (Embreston & Reese, 2000). These choices in the construction of the instrument are made in the pursuit of quality measures of the domain or trait, which also differ in personality research.

The trait.

The underlying trait being measured in personality research is often different in nature than an ability that might be measured in educational research. The usual assumption is that abilities follow an underlying normal distribution in the population. Hence, we expect to see a few geniuses and a few students with low abilities, but most of the population would be somewhere in between. However, traits such as self-esteem, as measured by a self-esteem scale targeted at distinguishing between low levels of self-esteem, may not follow a normal distribution (Embreston & Reese, 2000). Self-esteem may be bi-modal, with one group representing those with low-self esteem and the other group representing normal self-esteem. The nature of the trait, as measured by the particular instrument, should also be considered when applying IRT to personality research and in interpreting the results.

Much work and consideration has been given to employing both longitudinal models and IRT models to personality research. Using the lessons learned from conducting personality research, from each of these methodologies, this study will demonstrate the application of LG-IRM to personality research. Careful consideration will be given to modeling the response options and interpreting changes in self-esteem distribution over time. Using an example data set, it is shown how the IRT framework can provide interpretable information about how black self-esteem changes over time.

Method

The following details the study in which the LG-IRM is applied to a self-esteem measure included in the NSBA. A discussion of the particulars of the sample and instrument are included. A formulation of the LG-IRM for polytomous data is also presented.

Data.

NSBA data were used in this study, which included four waves. In the baseline year (1979-1980), the sample was selected as a nationally representative group of African-American households. Wave II took place in 1987-1988. Wave III took place in 1988-1989. Wave IV took place in 1989-1990. Waves I through IV are referred to as Waves 0 through 3 to be consistent with the LG-IRM parameterization. The NSBA was designed to provide information on physical and mental health, self-esteem, and life satisfaction among other variables to provide a basis for research on African Americans. *Self-esteem* can be defined in a number of ways including "level of global regard one has for the self" (Harter, 1993), or how well a person "prizes, values, approves, or likes" him or herself (Blascovich & Tomaka, 1991). *Self-esteem,* as measured in the NSBA, is consistent with both of these definitions.

Instrument.

The NSBA includes six self-esteem items derived from a combination of other self-esteem scales. The text and response options for each item are shown in Appendix C. Item 1, "I am a useful person to have around", was included to measure the worth aspects of self-esteem . Item 2, "I feel that I am a person of worth", and Item 5, "I feel that I do not have much to be proud of", come from the Rosenberg Self-Esteem scale (1965). Item 3, "I feel that I cannot do anything right", and Item 4, "I feel that my life is not very useful", were taken from an inventory

Bachman and Johnson used in the Monitoring the Future Project (Bachman & Johnson, 1978). Item 6, "As a person I do good job these days", was also added to measure the worth aspects of self-esteem. For each of the items, respondents were given the choice to answer in one of four response categories: "Almost Always True"; "Often True"; "Not Often True"; and " Never True". These response categories were scored in descending order from 1 to 4. To calculate the score for low-self esteem, the NSBA specified that the scores for items 3, 4, and 5 should first be reverse coded. The scores from each item should then be summed. Finally, 5 points should be subtracted from the total score, so that the total scores ranges from 0 to 15, with 15 representing low self-esteem and 0 representing the opposite. The reported reliability of thesix-item scale is 0.66 (Hughes & Demo, 1989).

Procedure.

To begin, the data were prepared for analysis using the Item Response Modeling software, ConQuest (Adams, Wu, & Wilson, 2005). As described above, this process includes recoding the negatively worded items and recoding of all of the items so that the lowest response level is zero (Appendix C). A process of model building was then started to determine how best to model the polytomous data and participation patterns of respondents cross-sectionally. The logic of including this model-building stage is that starting with the best-fitting cross-sectional model will produce a better-fitting longitudinal model. Several choices are available for modeling polytomous data. Among them is the Rating Scale Model the Partial Credit Model (Masters, 1981). The Rating Scale Model is actually a constrained version of the Partial Credit Model in which the step locations for each item are the same. Therefore, these models are nested and can be compared using the χ^2 distribution, where the degrees of freedom is equal to the difference in the number of parameters estimated for each model. To complete this test, separate models for self-esteem at each time point were first fit using the Rating Scale Model. The results from the Rating Scale Model showed consistent misfit of the step parameters for all waves of results. Misfit was defined as infit or outfit values beyond the range of 0.75 to 1.33. The results from the rating scale estimation suggest that the steps may not be the same for all items. The Partial Credit Model is an item Response model where the steps for each item are modeled separately, so it is likely to provide better fit. Next a Partial Credit Model was fit. The step fit statistics are shown in Table 12. The Unweighted Mean Square (UMNSQ), a measure of item parameter fit, showed some poor fit values for several item steps. For Step 3 the misfit was consistent across multiple waves. For instance, in waves 2 and 3 Item 1-Step 3 was misfitting. Also in waves 0 and 1, Item 3-Step 3 was misfitting. Furthermore, in all waves, Item 5-Step 3 was misfitting. However, when the infit (weighted MNSQ) is examined, these patterns do not appear. The infit weights the MNSQ so that a few discrepant cases do not overwhelm the results. All of the fit values are within the acceptable range. This suggests that the poor fit values may have been due to a few discrepant cases. When compared with the Rating Scale Model, which showed both poor infit and outfit statistics, the overall model also fits better for every wave. The comparison based on the χ^2 (Table 11) shows significance in every wave, suggesting that the increase in parameters associated with modeling each Item-Step combination separately improves the final deviance significantly. Thus, it was decided that the Partial Credit Model best models the response options of the self-esteem instrument included in the NSBA.

Table 11: Comparison of Models for Polytomous Data

Statistic	Wave	Rating Scale	Partial Credit***	Latent Regression Partial Credit
Final Deviance	0	21,013.6233	20,305.1587	20941.35072
df	0	9	19	20
Final Deviance	1	10,773.2822	10,461.7992	10736.11804
df	1	9	19	20
Final Deviance	2	8,068.5696	7,876.8571	8119.53882
df	2	9	19	20
Final Deviance	3	6,721.2991	6,523.3576	6726.13669
df	3	9	19	20

Table 12: Step Fit Statistics

Wave 0		Unweighted Fit				Weighted Fit				Wave 1		Unweighted Fit				Weighted Fit			
Item	Step	MNSQ	CI		T	MNSQ	CI		T	Item	Step	MNSQ	CI		T	MNSQ	CI		T
1	0	1.02	0.94	1.06	0.6	1.03	0.96	1.04	1.3	1	0	1.03	0.91	1.09	0.6	1.02	0.95	1.05	1
1	1	1.01	0.94	1.06	0.2	1.02	0.96	1.04	1.1	1	1	1.01	0.91	1.09	0.3	1.02	0.96	1.04	0.8
1	2	1.04	0.94	1.06	1.3	1	0.73	1.27	0	1	2	0.74*	0.91	1.09	-6.3	0.98	0.62	1.38	-0.1
1	3	1.83**	0.94	1.06	21.6	1.2	0.31	1.69	0.7	1	3	0.83	0.91	1.09	-4	1.04	0.37	1.63	0.2
2	0	0.96	0.94	1.06	-1.3	0.99	0.96	1.04	-0.6	2	0	1.01	0.91	1.09	0.3	1.02	0.95	1.05	0.7
2	1	0.97	0.94	1.06	-1.1	1	0.96	1.04	-0.1	2	1	0.99	0.91	1.09	-0.3	1	0.95	1.05	-0.1
2	2	0.62*	0.94	1.06	14.4	0.97	0.73	1.27	-0.2	2	2	0.95	0.91	1.09	-1	0.98	0.64	1.36	0
2	3	0.8	0.94	1.06	-6.9	1.19	0.26	1.74	0.6	2	3	2.39**	0.91	1.09	21.8	1.08	0.58	1.42	0.4
3	0	1.02	0.94	1.06	0.6	1.01	0.96	1.04	0.3	3	0	1.03	0.91	1.09	0.7	1	0.95	1.05	0.1
3	1	0.99	0.94	1.06	-0.4	0.99	0.97	1.03	-0.7	3	1	0.98	0.91	1.09	-0.4	0.98	0.97	1.03	-1
3	2	1.08	0.94	1.06	2.6	0.99	0.9	1.1	-0.2	3	2	1.01	0.91	1.09	0.2	0.97	0.84	1.16	-0.4
3	3	2.35**	0.94	1.06	31.9	1.08	0.8	1.2	0.8	3	3	1.45**	0.91	1.09	8.4	1.02	0.8	1.2	0.2
4	0	0.87	0.94	1.06	-4.4	0.93	0.95	1.05	-3	4	0	0.89	0.91	1.09	-2.4	0.93	0.94	1.06	-2.5
4	1	0.9	0.94	1.06	-3.4	0.96	0.94	1.06	-1.1	4	1	0.94	0.91	1.09	-1.2	0.98	0.92	1.08	-0.6
4	2	0.91	0.94	1.06	-2.8	0.95	0.89	1.11	-0.8	4	2	0.72*	0.91	1.09	-6.6	0.94	0.83	1.17	-0.8
4	3	2.3**	0.94	1.06	31	1.08	0.84	1.16	1	4	3	1.31	0.91	1.09	6.2	1.03	0.83	1.17	0.3
5	0	0.81	0.94	1.06	-6.7	0.9	0.94	1.06	-3.7	5	0	0.86	0.91	1.09	-3.2	0.9	0.94	1.06	-3.1
5	1	0.84	0.94	1.06	-5.4	0.95	0.93	1.07	-1.3	5	1	0.91	0.91	1.09	-2	0.97	0.9	1.1	-0.6
5	2	0.89	0.94	1.06	-3.8	0.96	0.85	1.15	-0.5	5	2	0.92	0.91	1.09	-1.8	0.95	0.84	1.16	-0.6
5	3	1.95**	0.94	1.06	24.1	1.09	0.85	1.15	1.1	5	3	1.71**	0.91	1.09	12.6	1.02	0.85	1.15	0.2
6	0	1	0.94	1.06	0.1	1.01	0.96	1.04	0.6	6	0	1.06	0.91	1.09	1.3	1.05	0.95	1.05	2.1
6	1	0.98	0.94	1.06	-0.7	1	0.97	1.03	0.1	6	1	1.01	0.91	1.09	0.3	1.02	0.96	1.04	0.9
6	2	0.79	0.94	1.06	-7.3	0.98	0.78	1.22	-0.2	6	2	0.89	0.91	1.09	-2.6	0.98	0.72	1.28	-0.1
6	3	6.59**	0.94	1.06	84.8	1.19	0.52	1.48	0.8	6	3	2.15**	0.91	1.09	18.8	1.06	0.6	1.4	0.3

Note. *MNSQ < 0.75, **MNSQ > 1.33

Table 12(cont.): Step Fit Statistics

Wave 2		Unweighted Fit				Weighted Fit				Wave 3		Unweighted Fit				Weighted Fit			
Item	Step	MNSQ	CI		T	MNSQ	CI		T	Item	Step	MNSQ	CI		T	MNSQ	CI		T
1	0	1.1	0.9	1.1	2	1.08	0.94	1.06	2.4	1	0	1.03	0.89	1.11	0.6	1.03	0.93	1.07	0.8
1	1	1.04	0.9	1.1	0.8	1.05	0.95	1.05	1.6	1	1	1	0.89	1.11	0	1.01	0.94	1.06	0.4
1	2	0.72*	0.9	1.1	-6.2	0.97	0.55	1.45	-0.1	1	2	0.78	0.89	1.11	-4.4	1	0.47	1.53	0.1
1	3	3.38**	0.9	1.1	29.8	1.02	0.44	1.56	0.1	1	3	1.92**	0.89	1.11	13.2	1.13	0.24	1.76	0.4
2	0	0.96	0.9	1.1	-0.7	0.99	0.93	1.07	-0.4	2	0	0.99	0.89	1.11	-0.2	1	0.93	1.07	0.1
2	1	0.94	0.9	1.1	-1.3	0.97	0.94	1.06	-0.9	2	1	0.96	0.89	1.11	-0.8	0.98	0.94	1.06	-0.5
2	2	1.11	0.9	1.1	2.1	0.95	0.56	1.44	-0.1	2	2	0.6	0.89	1.11	-8.4	0.96	0.47	1.53	-0.1
2	3	2.74**	0.9	1.1	23.7	1.04	0.41	1.59	0.2	2	3	9.21**	0.89	1.11	59.1	1.24	0.6	1.4	1.1
3	0	0.99	0.9	1.1	-0.1	0.98	0.93	1.07	-0.6	3	0	1.01	0.89	1.11	0.1	1	0.93	1.07	-0.1
3	1	0.96	0.9	1.1	-0.7	0.98	0.95	1.05	-1.1	3	1	1.01	0.89	1.11	0.1	1	0.94	1.06	-0.1
3	2	0.92	0.9	1.1	-1.5	0.95	0.81	1.19	-0.5	3	2	1.16	0.89	1.11	2.8	0.95	0.8	1.2	-0.5
3	3	0.99	0.9	1.1	-0.2	1.01	0.67	1.33	0.1	3	3	1.83**	0.89	1.11	12.1	1.09	0.64	1.36	0.6
4	0	0.85	0.9	1.1	-3.1	0.9	0.92	1.08	-2.6	4	0	0.78	0.89	1.11	-4.2	0.88	0.91	1.09	-2.9
4	1	0.88	0.9	1.1	-2.4	0.95	0.92	1.08	-1.2	4	1	0.84	0.89	1.11	-3.1	0.94	0.9	1.1	-1.2
4	2	0.58*	0.9	1.1	-9.7	0.9	0.78	1.22	-0.9	4	2	0.69	0.89	1.11	-6.4	0.95	0.74	1.26	-0.4
4	3	1.96**	0.9	1.1	15	1.08	0.74	1.26	0.6	4	3	1.29	0.89	1.11	4.8	1.03	0.72	1.28	0.2
5	0	0.83	0.9	1.1	-3.5	0.9	0.92	1.08	-2.6	5	0	0.81	0.89	1.11	-3.7	0.89	0.91	1.09	-2.5
5	1	0.86	0.9	1.1	-2.9	0.95	0.91	1.09	-1.1	5	1	0.83	0.89	1.11	-3.2	0.94	0.89	1.11	-1.1
5	2	0.83	0.9	1.1	-3.5	0.93	0.8	1.2	-0.7	5	2	1.52**	0.89	1.11	8.2	0.96	0.75	1.25	-0.2
5	3	1.66**	0.9	1.1	11	1.05	0.77	1.23	0.4	5	3	2.73**	0.89	1.11	21.6	1.07	0.74	1.26	0.5
6	0	1.08	0.9	1.1	1.6	1.06	0.93	1.07	1.9	6	0	1.06	0.89	1.11	1	1.04	0.93	1.07	1.2
6	1	1.05	0.9	1.1	0.9	1.04	0.94	1.06	1.4	6	1	1	0.89	1.11	0.1	1.01	0.94	1.06	0.5
6	2	0.63*	0.9	1.1	-8.4	0.95	0.59	1.41	-0.2	6	2	1.2	0.89	1.11	3.4	1.01	0.62	1.38	0.1
6	3	1.02	0.9	1.1	0.5	1.08	0.36	1.64	0.3	6	3	4.22*	0.89	1.11	33.4	1.14	0.3	1.7	0.5

Note. *MNSQ < 0.75, **MNSQ > 1.33

The participation patterns in the data also merit investigation. First, it is clear to see that the sample sizes decrease over time. In addition, it was found that the correlation between the number of waves and the estimated low self-esteem at the baseline, -0.0647, is significant (p=0.003). In applications of longitudinal models to educational achievement, it is important to determine whether the participation patterns are related to the construct being measured. For instance, if mathematics ability is being assessed, then ignoring students who drop out of school (and thus the sample) may induce bias, because their dropping out could be related to their mathematical ability. The importance of examining the missingness pattern is true in personality research as well. Therefore, it was important to determine if the participation pattern of respondents was related to their self-esteem. To evaluate this hypothesis, the number of waves of participation was calculated for each respondent. This number could range from 1 to 4. This number was then included in a latent regression with self-esteem for each time point (Appendix D). The results of the latent regression help to determine whether the participation patterns should be included explicitly in the model. The size of the regression coefficients show the amount by with the trait score should increase if another wave of participation were to be observed for that respondent. These coefficients are presented in Table 13.

Table 13: Results of Latent Regression

	Constant	SE	Participation	SE
Wave 0	-1.796	(0.041)	-0.073	(0.016)
Wave 1	-0.655	(0.117)	-0.182	(0.033)
Wave 2	-1.362	(0.309)	-0.122	(0.081)
Wave 3	-1.824	(0.556)	0.020	(0.144)

These findings show that a small increase in low self-esteem is found for participation in additional waves. This increase is quite small. The amount of increase is less than 0.2 logits for all waves. In addition, a comparison between the Latent Regression Partial Credit Model and the Partial Credit Model demonstrate that the Latent Regression Partial Credit model does not fit significantly better than the Partial Credit Model without the latent regression term at any time point (Table 12). Finally, the size of the correlation reported above, 0.0647, should be considered. Although significant (possibly because of the large sample), this correlation is near zero. Based on these findings, the Latent Regression Partial Credit Model with participation as a predictor was not included in the final cross-sectional model. Instead, it was decided that the Partial Credit Model was best cross-sectional model. These model-building sages provide evidence on how the LG-IRM should be parameterized for this data.

The LG-IRM.

With the growing popularity of latent growth modeling, measurement researchers have started to illustrate how latent growth can be represented as a multidimensional model. The models in this area posit that a person's responses at a certain time point are related to the person's latent variable at each time point. For illustrative purposes, the latent variables are thought of as factors. Each of the models described are illustrated in Figure 7 with three time

points. Andersen's (1985) model is a straight-forward example of this formulation. In this model, a separate factor is estimated for each time point. In Figure 7, latent variables are represented as circles and item responses are represented as squares. Thus the responses at Time 1 (X_{i1}) are related to the latent trait at Time 1 (ξ_1), and so on. Furthermore, to incorporate the measurement model, the response to Item 1 at all time points (X_{1k}) is related to that item's time-invariant

Andersen's Model Embretson's Model LG-IRM Model

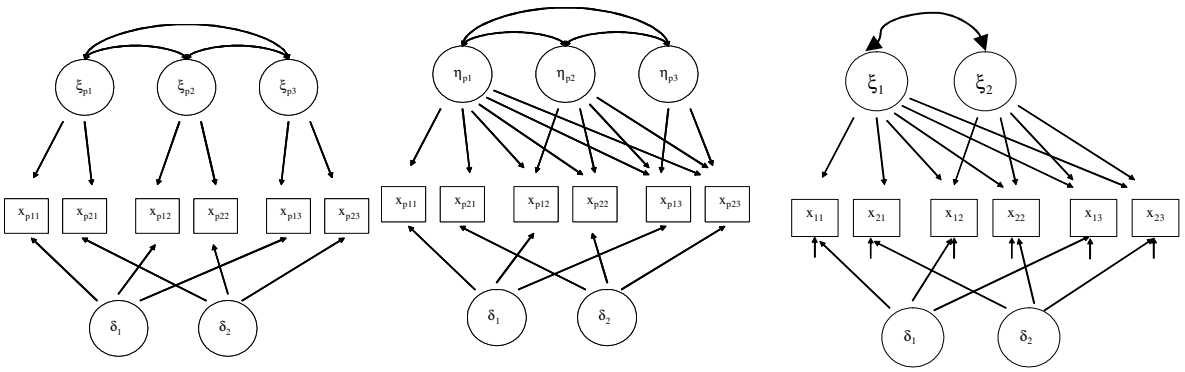


Figure 7: Growth model formulations
x_{ik} = response to item i at time k. ξ_j = latent trait (For Andersen and LG-IRM formulations). η_j = latent trait j (For Embretson formulation). δ_i = difficulty of item i.

difficulty, δ_1. Since the Andersen model specifies three factors, it allows for the difference between time points to be non-uniform, but it does not provide an overall growth estimate. In contrast, Embretson's (1991) formulation allows for an initial latent trait factor and then a growth factor between each consecutive set of measurement occasions. This formulation specifies one baseline parameter and two growth parameters as illustrated above. However, if only one linear growth parameter is needed for practical purposes, a simplifying assumption can be made. For instance, the growth between each successive time point can be constrained to be equal. In this more constrained model, both a baseline latent trait and single growth parameter are estimated. This is the formulation of the LG-IRM model (Wilson, Zheng, & Walker, 2007). The graphical formulation of this model illustrates how each of the item difficulties is related to the ability or latent trait over time in the case of dichotomous data. Yet, the more detailed explanation below, will illustrate how this model can be expanded to include polytomous data.

The Multidimensional Random Coefficients Multinomial Logistic model (MRCML) is a flexible multidimensional model that can be formulated to include polytomous items (Adams, Wilson, & Wang 1997; Briggs & Wilson, 2003; Wang, 1999). Thus, it is employed for estimation of the Latent Growth Item Response Model in the self-esteem example.

$$\mathrm{P}(X_{ip} = j) = \frac{\exp(\mathbf{b}_i' \vartheta_p + \mathbf{a}_{ij}' \xi_i)}{\sum_{k=1}^{K_i} \exp(\mathbf{b}_i' \vartheta_p + \mathbf{a}_{ik}' \xi_i)} \qquad (3)$$

Here, ϑ_p is a vector of latent abilities or traits. In the self-esteem example, this vector contains two elements; the level of low self-esteem at baseline (θ_{pb}) and the amount of increase in self-esteem from one wave to the next (θ_{pg}) where

$$\theta_{pt} = \theta_{pb} + t\theta_{pg} \qquad (4)$$

In The MRCML, the scores associated with selecting category j of each item, i, are specified using scoring vectors, $\mathbf{b_i}$. The collection of scoring vectors for each dimension is called the scoring matrix, denoted **B**. If the response to item *i* is scored on more than one dimension, this is called a *within-item multidimensional* model. The linear growth specification of the LG-IRM is a *within-item multidimensional* model because scores at time 0 (t=0) are only scored on the baseline dimension θ_{pb}. But all scores thereafter are scored on both the baseline and growth dimensions(θ_{pb}, θ_{pg}). In ConQuest, the scoring matrix is defined in the command syntax (See appendix D). For models with polytomous data, $\boldsymbol{\delta}$ is a vector of item difficulties and steps. The design vector $\mathbf{a}_{ij}$ is used to specify the combination of the item and step parameters corresponding to a response in any category $X_{ip} = j$. In the case of the Partial Credit Model, where each item has the same number of response categories, $\boldsymbol{\xi}$, is a vector of *n* item difficulties, and *n(k-1)* step parameters, where k is the number of response categories $\boldsymbol{\xi}$=[$\delta_1, \delta_2, \ldots, \delta_N, \tau_{11}, \tau_{12}, \ldots, \tau_{1(K-1)}, \ldots \tau_{N1}, \tau_{N2}, \ldots, \tau_{N(K-1)}$]. A design vector a_{ij}, is then used to specify the pararametes of each δ_i that should be used in modeling $X_{ip} = j$. The collection of these individual design vectors is called the *design matrix* and denoted A. When modeling longitudinal data, the response vector will include responses from each wave of administration. Thus, the design matrix must define the combinations of item and step parameters involved in modeling the responses to the item during a particular wave $X_{ipt} = j$. In the case of the LG-IRM, the same item and step parameters are used in modeling the response $X_{ipt} = j$ at all time points. Hence, the design matrix is essential expanded for each point in time. An example of the design matrix construction for the LG-IRM is shown in Figure 8. The complete design matrix and the design block for one year are shown. Only one design block, for time 1, is highlighted because the other are essentially replications of that block. The design block contains 17 columns and 24 rows. The design block shows the parameterization of the 5 items and the 2 steps for each item, resulting in a total of 17. Item 6 and Step 3 for each item are constrained. With the addition of these parameters, the total is 24. By using a combination of the expansion of the design matrix for one point and the scoring matrix, the longitudinal data can be modeled using the MRCML framework.

Results

The LG-IRM was applied to the self-esteem data. The results show that averaged over the years from Grade 7 to Grade 12, low self-esteem tends to increase (Table 14). The mean baseline ability ($Mean_b$) is -1.5930 logits. It increases by 0.0640 logits in each year ($Mean_g$). This increase was estimated on the logit scale and can therefore be compared to the location of the items. A graphic (Figure 9) was produced to illustrate this quality. Figure 9 shows an Annotated Wright Map. The distribution of self-esteem is shown at baseline (Wave 0). The distribution of self-esteem is also shown at the point of the final administration (Wave 3). The construct measured was low self-esteem (to be consistent with the NSBA) so larger positive values represent low self-esteem and larger negative values represent better self-esteem. The Annotated Wright Map

Figure 8: Design matrix for administration over four years

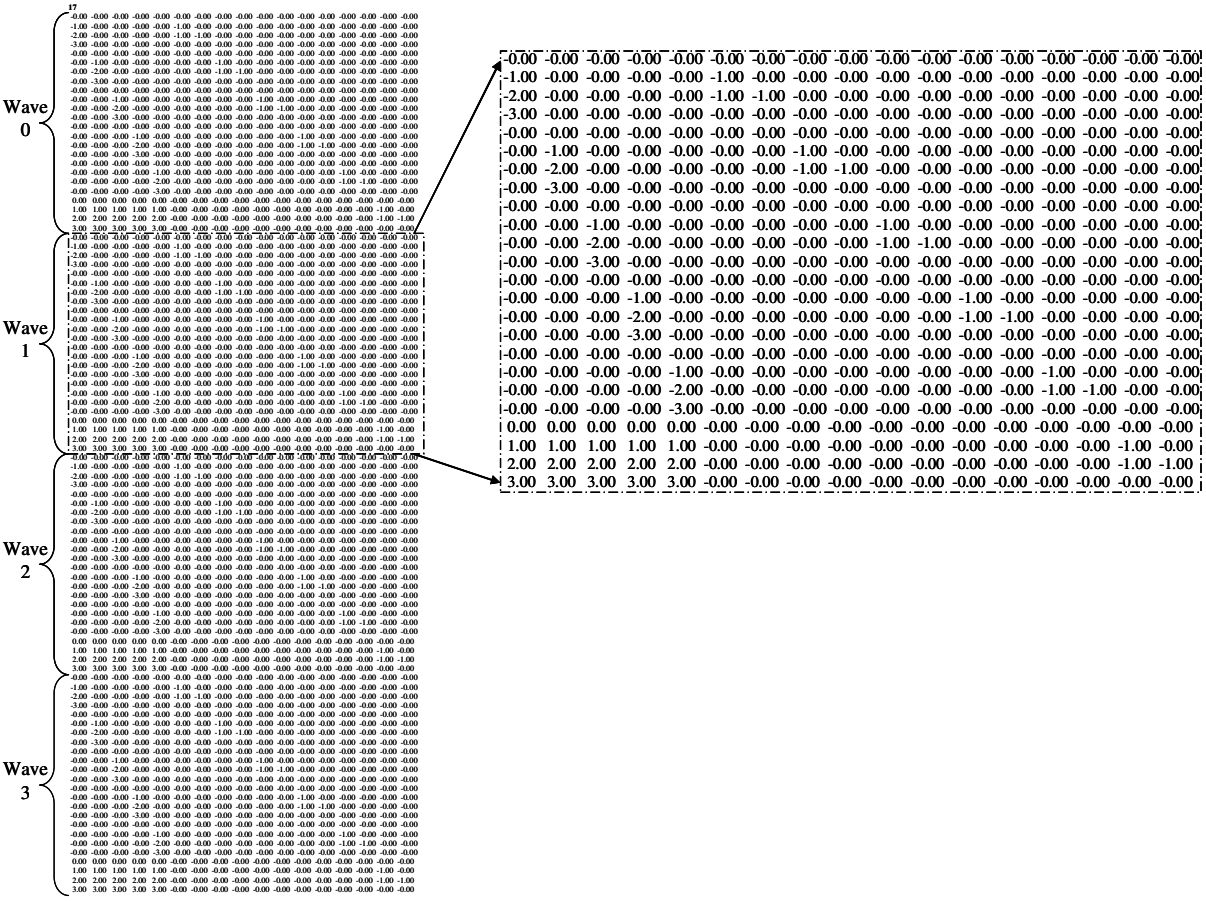


Figure 9: Annotated Wright Map showing weighted likelihood estimates and item difficulties

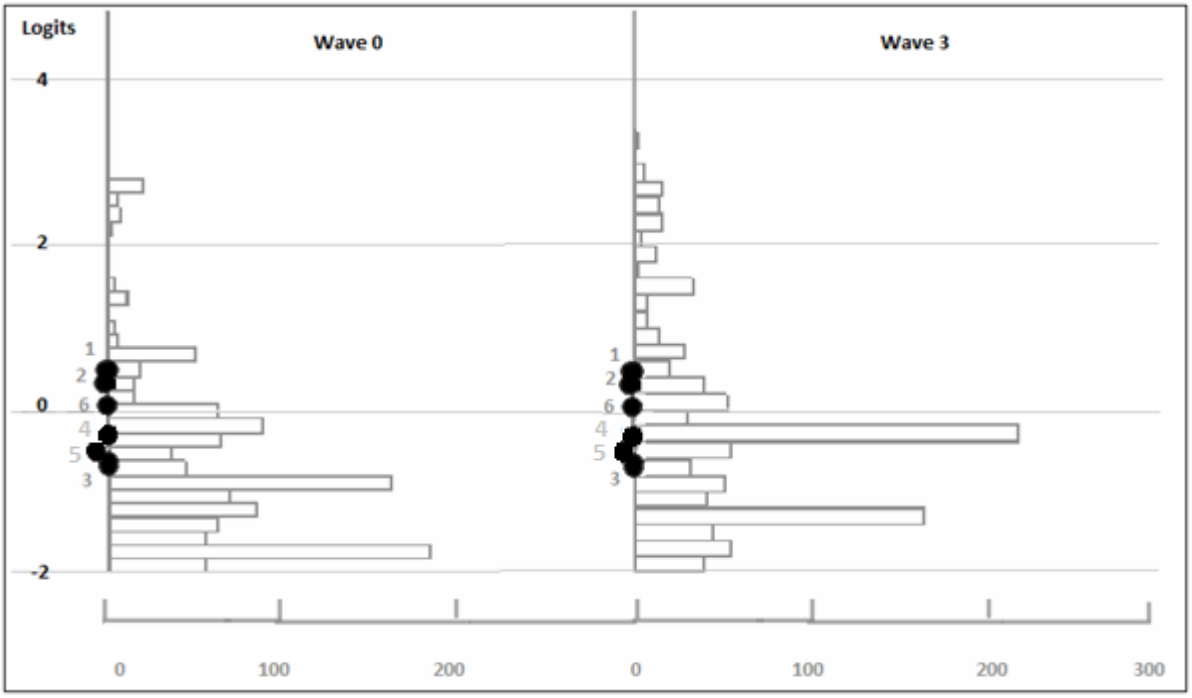


Table 14: Step Fit Statistics From Cross-sectional Partial Credit Models

			Unweighted Fit				Weighted Fit			
Variable	Estimate	SE	MNSQ	CI		T	MNSQ	CI		T
$Mean_b$	-1.5930	0.0210								
$Mean_g$	0.0640	0.0120								
$Variance_b$	0.9080									
$Variance_g$	0.2950									
$Correlation_{bg}$	-0.2950									
Item 1	0.3788	0.0194	0.88	0.94	1.06	-3.9	0.94	0.92	1.08	-1.4
Item 2	0.2786	0.0191	0.86	0.94	1.06	-4.9	0.94	0.92	1.08	-1.6
Item 3	-0.3254	0.0166	1.16	0.94	1.06	5	1.3	0.92	1.08	6.8
Item 4	-0.2798	0.0162	1.18	0.94	1.06	5.6	1.28	0.92	1.08	6.2
Item 5	-0.2888	0.0161	1.16	0.94	1.06	4.9	1.28	0.92	1.08	6.1
Item 6*	0.0473									
Item 1 Step 1	-1.4956	0.0330	1.16	0.94	1.06	5	1.36	0.94	1.06	10
Item 1 Step 2	1.4925	0.1042	1.01	0.94	1.06	0.2	1.23	0.8	1.2	2.1
Item 1 Step 3*	0.0031									
Item 2 Step 1	-1.1115	0.0338	1.05	0.94	1.06	1.5	1.17	0.94	1.06	4.9
Item 2 Step 2	1.4277	0.1042	0.59	0.94	1.06	-15.7	1.18	0.8	1.2	1.7
Item 2 Step 3*	-0.3161									
Item 3 Step 1	-0.9058	0.0324	1.14	0.94	1.06	4.3	1.28	0.94	1.06	8.1
Item 3 Step 2	0.4871	0.0530	1.09	0.94	1.06	3	1.08	0.9	1.1	1.5
Item 3 Step 3*	0.4186									
Item 4 Step 1	-0.0579	0.0350	0.99	0.94	1.06	-0.4	1.12	0.93	1.07	3.5
Item 4 Step 2	0.1519	0.0581	0.92	0.94	1.06	-2.7	1.14	0.89	1.11	2.3
Item 4 Step 3	-0.0940									
Item 5 Step 1	0.2199	0.0367	0.95	0.94	1.06	-1.8	1.13	0.93	1.07	3.4
Item 5 Step 2	0.1261	0.0619	0.84	0.94	1.06	-5.6	1.13	0.88	1.12	2.1
Item 5 Step 3*	-0.3460									
Item 6 Step 1	-1.2511	0.0332	1.12	0.94	1.06	3.6	1.25	0.94	1.06	7
Item 6 Step 2	1.1769	0.0876	0.88	0.94	1.06	-4	1.26	0.83	1.17	2.8
Item 6 Step 3*	0.0742									

shows that the overall distribution shifts up slightly from Wave 0 to Wave 3. This shift represents the slight increase in low self-esteem mirrored by the linear growth estimate (0.0640 logits/year). The annotated Wright Map also shows the locations of the items. For ease of representation, the overall item difficulties are shown. The values plotted for Items 1-6 are the estimates of δ_1, δ_2, δ_3, δ_4, δ_5, and δ_6. These can be interpreted as the average difficulty of the items, and the step parameters τ_{ik} can be interpreted as deviances from the average difficulty. The estimates for the entire set of item parameters, $\boldsymbol{\xi}$, is shown in Table 14. The estimate for δ_1- δ_5 is listed as Item 1-5. The value for Item 6 was not actually estimated, but used to constrain the mean of the item distribution to zero. In addition, the estimates of the step parameters are listed for each item. Since the overall difficulty of the item is interpreted as the average difficulty of the item and the step parameters are deviances from this difficulty, the average of the step parameters is zero. Therefore the values for Steps 1 and 2 are estimated, but that for Step 3 is calculated so that the average of the step parameters is zero. The Weighted MNSQ values show that all of the items and step parameters fit well except for Item 1-Step 1. Item 1, " I am a useful person to have around," is an item that was added to the Self-Esteem inventory by the NSBA creators. The step parameter represents the step between the response options "*Almost Always True (0)"* and *"Often True (1)."* This item was also the most difficult on the low self-esteem scale. Hence, this item would be associated with the lowest levels of self-esteem. It could be possible that the misfit is due to its positioning on the scale. Self-esteem scales are designed to distinguish between respondents with low self-esteem, but it may be possible that the instrument does not work as well at the very low end of the scale.

Individual growth trajectories were also plotted to illustrate the variation in self-esteem trajectories over time (Figure 10). The correlation between the baseline level of the trait and the growth rate is negative (-0.259 logits/year). Thus, it should be expected that respondents with lower self-esteem are associated with larger improvements and that respondents with less low self-esteem are associated with less improvements. The trajectories shown in Figure 10 represent a random selection of from the NSBA respondents. The figure also illustrates that each person's baseline level of low self-esteem and growth rate are specific to that individual.

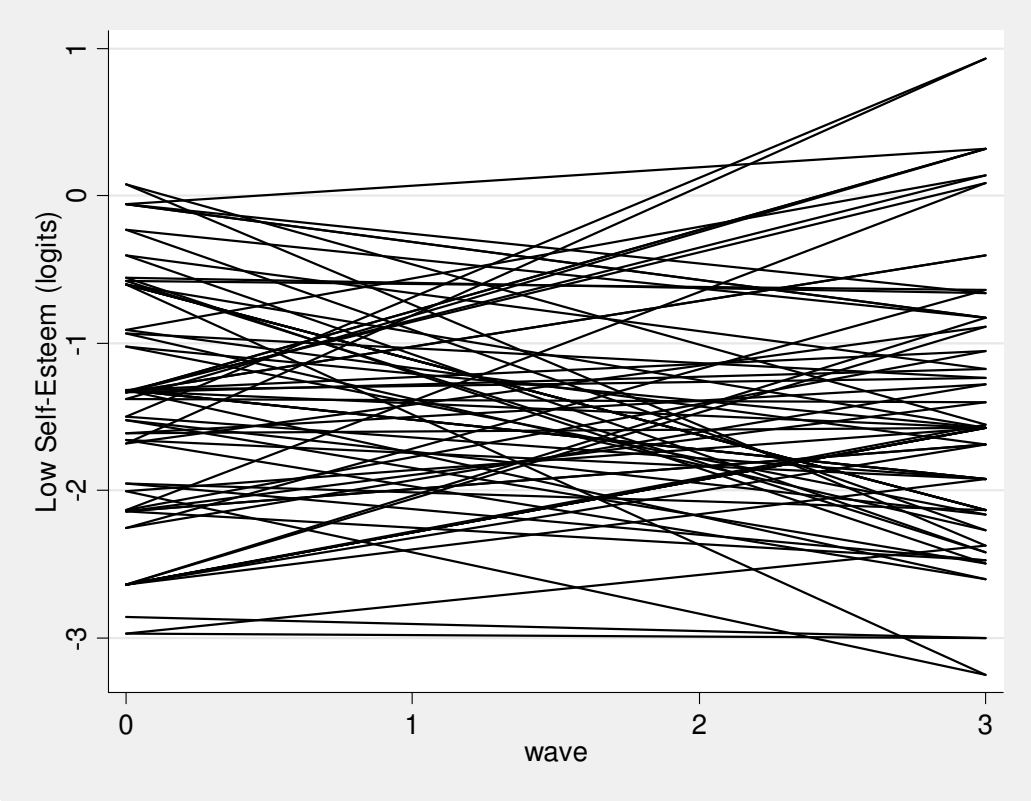


Figure 10: Individual Self-Esteem Linear Growth Trajectories. This plot shows a random sample of linear growth trajectories plotted using the WLEs for the baseline self-esteem and linear growth parameters from the NSBA example

Discussion

In this study, the LG-IRM was applied to a self-esteem inventory included in a national longitudinal survey of Black Americans. To begin building this model, cross-sectional item response models were fitted to the self-esteem inventory included on the NSBA. The results showed that the best fitting model for the Likert-type items was the Partial Credit Model. A partial credit formulation of the LG-IRM was then presented. In this presentation, the parameterization of the item difficulties and step parameters was discussed. Creation of the design matrix was also presented. Finally, the LG-IRM was estimated for data from four waves of the NSBA.

The results of the LG-IRM show that, on average, low self-esteem continues to decrease slightly over time. This finding may be consistent with the theory that self-esteem decreases in late age. The NSBA administrators did note that the survey included a slight oversampling of older adults (Jackson & Neighbors, 1996). Thus, the proportion of older adults in this sample may have influenced the results. Further study might include an analysis that includes the ages of the respondents in the model.

This study also presented tools for interpreting the amount of change in low self-esteem over time. The Wright Map (Figure 9) shows how the distribution of low self-esteem shifts over time relative to the item locations. At Wave 0, the largest group of respondents is located below the average difficulty of Item 3: "I can't do anything right." This item was associated with the lowest levels of low self-esteem. However, at the end of the survey administration (Wave 3), the largest group of respondents is located near the average difficulty of Item 4, "My life is not very useful," which was an item associated with a higher level of low self-esteem. In addition, the largest group at Wave 3 has also passed the level of Item 5: "I don't have much to be proud of." So the interpretation of the change over time could be that respondents are experiencing fewer feelings of pride and competence. These interpretations are based on the average difficulty because the the measurement was conducted on not an entirely clinical population. Therefore, relative to the entire span of the scale, the Item-Steps would be very close to the average item difficulties. However, if this method were to be used for data from a clinical population, whose scores could be expected to be more tightly distributed around the location of the items, it could be useful to include the step locations in the Wright Map to describe interpretations of smaller increments of change in self-esteem.

Another way to incorporate the values of the step parameters would be to interpret the total raw scores. An interpretation of the total raw score also facilitates the connection between the original scores, as calculated following the instructions of the NSBA. Using ConQuest, scores on the logit scale are written in terms of the raw score (Table 15). The raw scores, as calculated by ConQuest, include any recoding that has been done to prepare the data for calibration. Thus, so an additional column is shown to provide the mapping of the raw score to the original raw score from the NSBA. The correspondence in Table 15 could be used to interpret the scores. For instance, the range of average item difficulties is roughly from -0.3 to 0.4 logits. These logit values represent raw scores in the range from 7 to 11. They also encompass the range from 0 to 1 points on the original scale. Connections such as these could be made to illustrate how logit scores could be connected to the scoring system that may be more familiar to users of self-esteem inventories.

Table 15: Logit Raw Score Equivalence Table

Raw Score	SE Score	Logit Equivalence (WLE)	SE
0		-3.6972	1.5282
1		-2.4639	0.9082
2		-1.8342	0.7143
3		-1.4064	0.6099
4		-1.0868	0.5462
5		-0.8290	0.5064
6		-0.6059	0.4823
7		-0.4007	0.4695
8		-0.2012	0.4656
9	0	0.0020	0.4690
10	1	0.2169	0.4781
11	.	0.4487	0.4909
12	.	0.6970	0.5056
13	.	0.9566	0.5221
14		1.2262	0.5444
15		1.5145	0.5805
16		1.8463	0.6456
17		2.2856	0.7818
18	15	3.1336	1.2444

The results illustrate how IRT can provide meaningful interpretations of the level of self-esteem at a particular point in time, as well as the amount of change in self-esteem. Attempts were made to illustrate the utility of IRT methods for personality research using illustrations, plots, and score translation. Further efforts to translate the results from IRT models into a useable form for clinicians and personality researchers will only help to facilitate the incorporation of IRT into personality research.

Chapter 3: Exploring Cross-Cultural and Time-Wise Item Shift: An Extension of the Latent Growth Item Response Model

Exploring Cross-Cultural and Time-Wise Item Shift: An Extension of the Latent Growth Item Response Model

The No Child Left Behind Act (NCLB, 2002) stipulates that states implement accountability systems to measure progress. In many cases, the measurement of adequate yearly progress (AYP) is conducted by comparing scores of successive cohorts of students. To quantify individual growth, it is better to measure the same students over time. Many studies do not employ repeated measures designs due to lack of time or resources to take multiple measurements. However, states can apply for multi-million dollar grants to design and implement longitudinal data systems under the Educational Technical Assistance Act (NCLB, 2002). Since repeated measures designs present a considerable strain on fiscal resources in any research study, it is important to ensure that the measurements taken will be suitable for comparison over time. Progress cannot be quantified accurately unless the measurements are exact.

The process of combining common items and common persons is often used within psychometrics and is called equating (Holland & Rubin, 1985). Typically items that are common to an instrument across waves of data collection are considered to be comparable. The common items serve as a means of anchoring the waves of data to each other. The equating process is aimed at producing comparable scores for different measurements and different pupils. This process is usually conducted using measurement specific software and rarely discussed outside of psychometric circles. Many methods are available to test publishers to conduct vertical equating studies, each with its own set of assumptions. Discussion of all of the methods of vertical equating is beyond the scope of this paper (see Holland & Rubin, 1985). However, they all rely on the stability of the common items. In some cases, as will be demonstrated in this study with data from the first three waves of National Educational Longitudinal Study 1988 (NELS, 2000), these common items are not stable. This presents a problem for typical longitudinal modeling because the assumptions of the model are violated. This paper will explore how assumption violations affect the estimates and how models can adapt to assumption violations.

Another reason to examine the changing nature of constructs over time is that these shifts may occur differently in certain groups of students. For instance, males and females could experience the same mathematics items differently as they mature, possibly as gender stereotypes become more salient (Steele & Aronson, 1995). Although an item may be equally different for both genders, it will be more difficult for females over time. Similarly, students from different baseline proficiency classifications may experience the same items as being increasingly hard or increasingly easy as they progress through understanding of a certain topic. These differing patterns present another reason for exploring and attempting to model construct changes within the growth model.

This paper reviews the current work in both detection and incorporation of changes in constructs into growth models. I also present an example in which the construct does seem to change over time. In addition, these changes seem to be different within two groups of students. In order to estimate growth for these students, different models are considered. The cross-sectional models allow for accurate estimates of group differences and differences in the construct over time, but do not provide student growth parameters. Therefore, the use of an item response model for growth is demonstrated. In addition, an extension of this model is presented, which allows for differences in item parameters across time.

Longitudinal Measurement Invariance

Although this study examines the effect of changing constructs over time when using an item response model, the prevalence of research in this area has been conducted within the factor analytic framework. Referring to growth models specifically, the term *longitudinal measurement invariance* is defined as the comparability of measurements over time (Vandenberg & Lance, 2000). Studies from multiple application areas warn that a change in the amount of a latent trait or ability can only be interpreted as growth when longitudinal measurement invariance is ensured (Pentz & Chou, 1999; Rahu, Laffitte, & Byrne, 2002). Thus, researchers within the factor analytic framework are concerned with longitudinal measurement invariance for reasons similar to those interested in using item response theory for growth modeling.

The factor analytic framework also presents various means of ensuring longitudinal measurement equivalence using statistics readily available within this framework. Although this framework is widely applied, the derived factors are not generalizable to different groups and do not take into account the level of the latent trait involved (Lambert, Essau, Schmitt, & Samms-Vaughan, 2006). In addition, the constructs are examined as a whole, not at the item level. This approach does not provide any information on how specific items change in nature over time. Lambert and his colleagues also argued that item response modeling techniques provide information on the items that are working or are not working. Given the advantages of item response modeling, and the inapplicability of factor analytic techniques in this area, it is useful to look within the item response modeling framework for studies of longitudinal measurement invariance.

Few studies address the issue of item shift over time, within the item response modeling framework. Therefore, there are few guidelines for application directly to the Latent Growth Item Response Model (LG-IRM). In a recent study, Long, Harring, Brekke, Test, and Greenberg (2007) illustrated that longitudinal measurement invariance requires equal discrimination parameters over time and equal (i.e. invariant) location parameters. These researchers used a more complex form of item response function than the Rasch model on which the LG-IRM is based. In the Rasch model, the difficulties are already constrained to be the same across all items and time points, making the recommended constraint on discrimination parameters not directly applicable. Thus, it is necessary to look deeper into the literature for guidance in considering issues of longitudinal measurement invariance when employing the LG-IRM.

Other types of invariance.

Given the limited body of research on longitudinal measurement invariance using item response models, it is helpful to look to studies of other types of invariance. In their review of the literature in this area, Vandenberg and Lance (2000) agreed that the definition of *measurement invariance* is the degree to which measures are comparable across groups or time. This suggests that similar techniques and frameworks can be used to examine the changing nature of constructs over time or across groups. Within the item response modeling framework, constructs are often examined for differences across groups. Reise, Widman, and Pugh (1993) illustrated that to ensure measurement invariance the item difficulty parameters must be the same across groups. This is a classical approach that goes back as far as Wright and Stone (1979) and Lord (1981). Normally, groups are thought of as exclusive subsets of a sample as defined by some characteristic. Common examples are proficiency level, age, or gender. Another way of creating groups is to identify the time of data collection. In the case of longitudinal data, groups could be

created by dividing the sample into waves of data. Thus, techniques for examining measurement invariance could be applied to time points.

Incorporating violations of invariance.

The item response modeling framework also provides methods for incorporating violations of invariance into the model. One way to include group differences into the item response model is to include differential item function (DIF) terms. These additional parameters allow for the item parameters to be slightly different across groups for cross-sectional data. Their values represent the difference in difficulty for a certain item between two different groups. They can also provide a way to estimate differences in items across time for longitudinal data. In this case, the value of the DIF parameter represents the difference in an item parameter at each successive time point. Given that the DIF model provides a reasonable way to determine the amount of item shift over time, it is a good way to evaluate longitudinal measurement invariance. Significant DIF would provide evidence of violations of the invariance assumption. However, the DIF model does not provide growth estimates when applied to longitudinal data. To measure student progress over time it is still necessary to employ the LG-IRM.

The present study.

This study uses the LG-IRM with a data set in which item shifts do occur. Specifically, it shows how the LG-IRM can adapt to cases in which items are not constant over time. Two methods to quantify these shifts are used. In the first, the item parameters for each year are estimated separately and compared. In the second the item shifts are estimated using all three waves of data by incorporating a special parameter for the differences. To explore group differences, these models are estimated for two different subsets of the sample. By considering evidence of item shift when parameterizing the LG-IRM, the study shows how the model can adapt to violations of longitudinal measurement invariance.

Method

Data.

The data for this study were sampled from NELS. The history subtest has 30 items and covers topics in American History. The study includes data from the history test that was administered during three waves of the NELS (1988, 1990, & 1992). The set of common items that were administered in all three measurement occasions consists of 17 core items. Only the common items are used in the growth modeling in this paper because the LG-IRM has only been tested through simulation in these cases (Wilson, Zheng, & Walker, 2007). Once their estimability had been verified through simulation, data sets with both common and unique items could be calibrated.

The waves of NELS data used in this study consist of a nationally representative sample of students in Grades 8, 10, and 12. The analysis sample was restricted to those who participated in all three test administrations. The sample was restricted to students who participated in all three waves of data collection to simplify the setup of the longitudinal model. This resulted in a sample size of 11,552 students. Models that can incorporate missing data resulting from various processes require additional modeling considerations. One process that might result in missing data is self-selection. If students choose to be present on the day that they will take the NELS in each wave due to self-perceived prediction of success in such endeavors, it is hard to make

comparisons across years without complicated models. These sampling issues are beyond the scope of this paper.

In order to compare the performance of different groups over time, the sample was divided into two groups based on their performance on five additional items measuring cultural sensitivity. These items measure students' awareness of important historical events relevant to immigrant, minority, and native groups in the United States. Low scorers, or those who answered two or less of the five items correctly, were placed in the Reference Group (N_r=5,748). High scorers, or those who answered three or more of the five items correctly, where placed in the Culturally Sensitive Group (N_{cs}=5,804). These two groups provide the basis for comparison of growth patterns in learning history over time. In addition, the effect of item shifts over time can be compared across these two groups.

Procedure.

For two subsamples of students, the item and person parameters were estimated in four ways. Data sets were calibrated consecutively, once at each time point. The item parameters from this analysis are compared to show the differences in item parameters across time. The data sets were also calibrated using data from all three points in time, incorporating the time parameter as an item-by-year DIF term. This model incorporates three waves of data into the same model even though the item parameters differ with time. These cross-sectional approaches are used to demonstrate typical approaches to modeling this type of data and to present justification for moving to the more complex LG-IRM. The parameters were estimated using the LG-IRM with the constant item assumption. Given the DIF parameters estimated previously, the LG-IRM was expected to be slightly inaccurate. Finally, an LG-IRM model was estimated in which the item parameters are estimated freely. A comparison of these models demonstrates the item shifts over time and their effects on the LG-IRM model for each sample. Additionally, a comparison of the results from these models across the two subsamples of students demonstrates how these patterns might be different across groups. ConQuest (Wu, Adams, & Wilson, 1998) software was used for all types of estimation.

The model.

The LG-IRM is a unique model because it is formulated and estimated as an item response model. To calculate the person ability at any point in time θ_{pt}, the linear growth version of the LG-IRM specifies a relationship between the a person's ability at baseline ($t=0$) and future points in time ($t=1, 2, \ldots$) using a intercept (θ_{pb}) and a slope term (θ_{pg}).

$$\theta_{pt} = \theta_{pb} + t\theta_{pg} \tag{5}$$

Since a linear model is specified, ability at each individual time point does not need to be estimated. Only the baseline ability and growth needs to be estimated. The LG-IRM formulation (Wilson et. al., 2007) then shows how the probability of a correct response from person j to any item at time t (x_{it}), can be written as a function of the difference between person p's ability at time t (θ_{pt}) and the difficulty of item i (δ_i).

$$\mathrm{P}(x_{ipt}=1) = \mathrm{P}_{ijk} = \frac{\exp(\theta_{pt} - \delta_i)}{1+\exp(\theta_{pt} - \delta_i)} \tag{6}$$

The equations presented here are meant to communicate clearly the type of growth defined by the application of the LG-IRM in this study. There are, however, several other features of the model that deserve mentioning. In addition to the linear relationship specified in Equation 2, it should be noted that in both the simulation and examples in this paper, the baseline ability and growth are positively correlated with each other. In relation to multidimensional models such as the multidimensional three-parameter logistic model, the LG-IRM is conceptually different because the dimensions are not intended to be orthogonal. Also, to define the mapping of items onto each of the dimensions, a scoring matrix is used. This becomes especially important when different items are used in each wave of testing. Finally, a design matrix is used to define the linear combinations of the item parameters. In practical terms, a main purpose of the design matrix is to set the proper constraints for estimation. For the common items, it is used to constrain these items to have the same difficulty for each year of testing in which they are presented. The design matrix is also used to identify any constraints being place on the items to identify the model. Specifications of the scoring matrix and design matrix were discussed in earlier chapters.

Estimation.

The LG-IRM model can then be estimated with the MRCML model with two person dimensions. The first dimension is the baseline latent trait and the second is the growth factor. Furthermore, since the LG-IRM model is one of a generalized set of flexible models, the LG-IRM model can be expanded to include other models for non-linear growth, differential item function, and item bundles.

The LG-IRM framework presents an addition to the growth modeling framework with much potential so the utility and accuracy of this model must be verified. For instance, all of these growth models specify that the items have a constant difficulty over time. In terms of the notation presented in this study, each item only has one difficulty, δ_i, across time points. δ_i is not also indexed by time, *t*. This is the longitudinal measurement invariance assumption. Though this assumption is made in estimating these models, it may not always be true as will be shown in the NELS example.

Results

The results from the four estimations are described below. First, parameters for the Rasch model were estimated separately for the three estimation occasions of each sample. These results show that item parameters change over time and that these change patterns are similar across groups. The person parameter estimates also show growth patterns within groups. Second, parameters were estimated using an item-by-year DIF term for each sample. The item-by-year DIF provides further evidence as to how the items shift over time. Evidence from both of these cross-sectional models is discussed. Next, parameters were estimated using the LG-IRM model for each sample. Estimates from this model are interpreted carefully, given the evidence of item shift from the two simpler models. Finally, estimates from the LG-IRM model extension, which allows for item shifts, are presented.

Cross-sectional models.

Separate calibration.

In some ways, the item parameters seem reasonably similar across measurement occasions. This is shown graphically for the Full Sample in Figure 11 where the item parameter estimates for each of separate estimations is shown in scatter plots. It is clear that they all lie close to an invisible diagonal line that would represent perfect correlation. The same pattern exists for the Reference Group as shown in Figure 12. Figure 13 shows that the item difficulties for the Culturally Sensitive Group follow a similar pattern. The actual correlations between item estimates in each measurement occasion are high, as could be guessed from the scatter plots. For the Full Sample, the correlation between item estimates in the baseline year and first follow-up year, ρ_{01}= 0.964. The correlation between item estimates in the first follow-up and second follow-up, ρ_{12}=0.931, is also high. Finally, if the baseline estimates are correlated with the second follow-up estimates, this correlation, ρ_{02}= 0.900, is also high. The lower correlation between the first and final wave of data might be explained by the items shifting a bit more during each time interval. This would result in a lower correlation between the item estimates for the first wave and the final wave because they are separated by more time.

This pattern of correlations between items estimates, across the three waves of data, is the same for the Reference Group and the Culturally Sensitive Group as shown in Figures 12 and 13. The high correlations between items at each measurement occasion provide preliminary evidence that items are relatively similar over time. Thus, some might conclude that item shift is not problematic and that the assumption of longitudinal measurement invariance has not been violated. However, correlation results often hide more nuanced differences. Further examination of the differences reveals interesting patterns in how items actually shift.

Differences in the item difficulties were calculated to quantify how the items shift between measurement occasions under the separate calibrations schemes. These differences are shown in Table 16 for the Full Sample, Table 17 for the Reference Group, and Table 18 for the Culturally Sensitive Group. Interval 1 shows the difference between the estimates in the baseline and first follow-up year, and Interval 2 shows the difference between them in the first and second follow-up. Finally, the Overall difference shows the change from the baseline year to the second follow-up year. Differences larger than 0.5 logits on the history scale are emphasized in the tables with asterisks. For all three of the analysis samples, more items were identified as having large amounts of difficulty shift over Interval 2 as compared to Interval 1. This may mean that the period between the first and second follow-ups is very important to ensure the equivalence of the items over testing occasions. However, for all of the analysis samples, fewer items were identified as having large amounts of difficulty shift. This could mean that some of the larger item difficulty changes across particular intervals level out when the overall item difficulty shift is calculated.

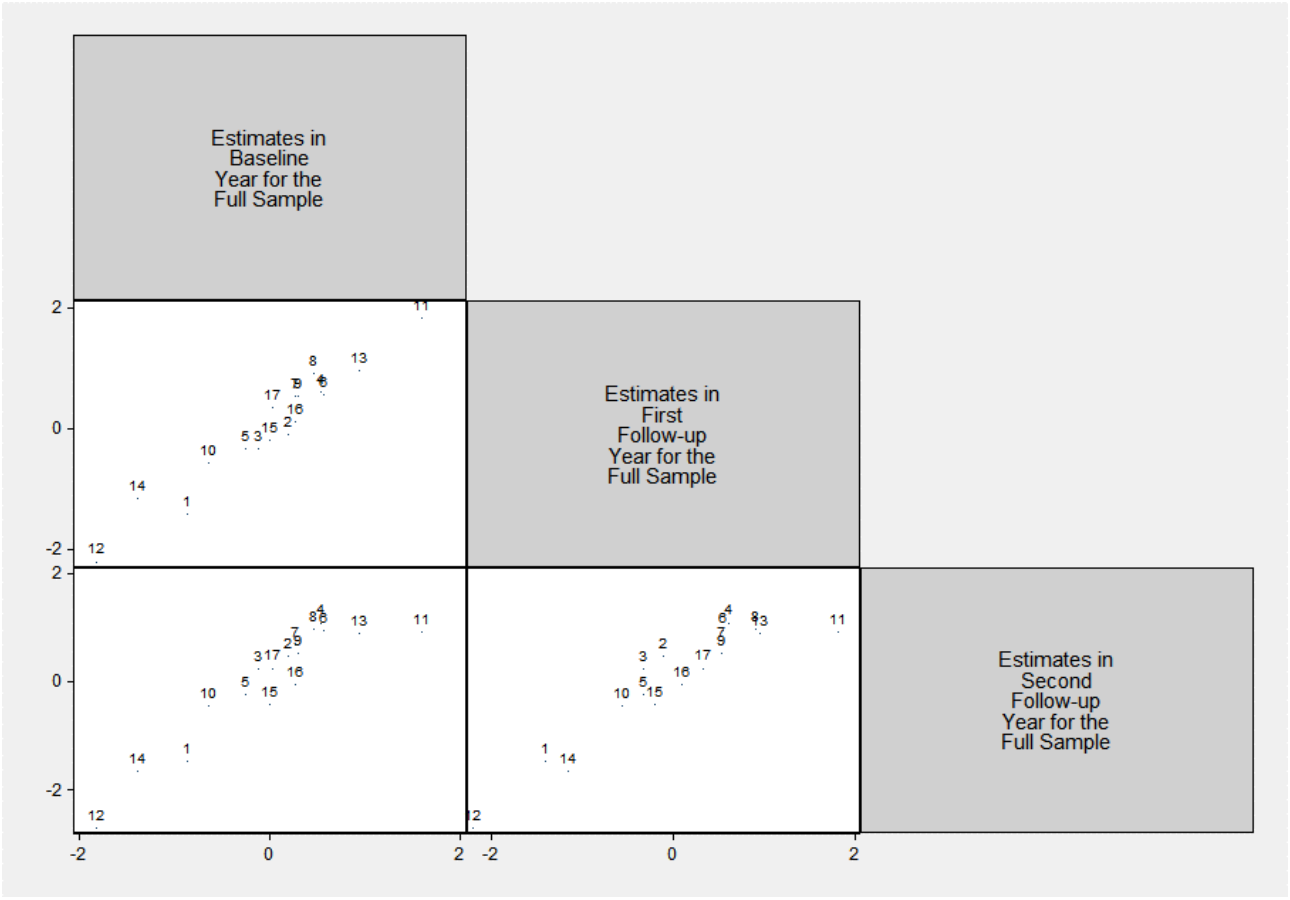


Figure 11: Relationship between item estimates for each year in the full sample. r_{01}= 0.964, r_{02}= 0.900, r_{12}=0.931

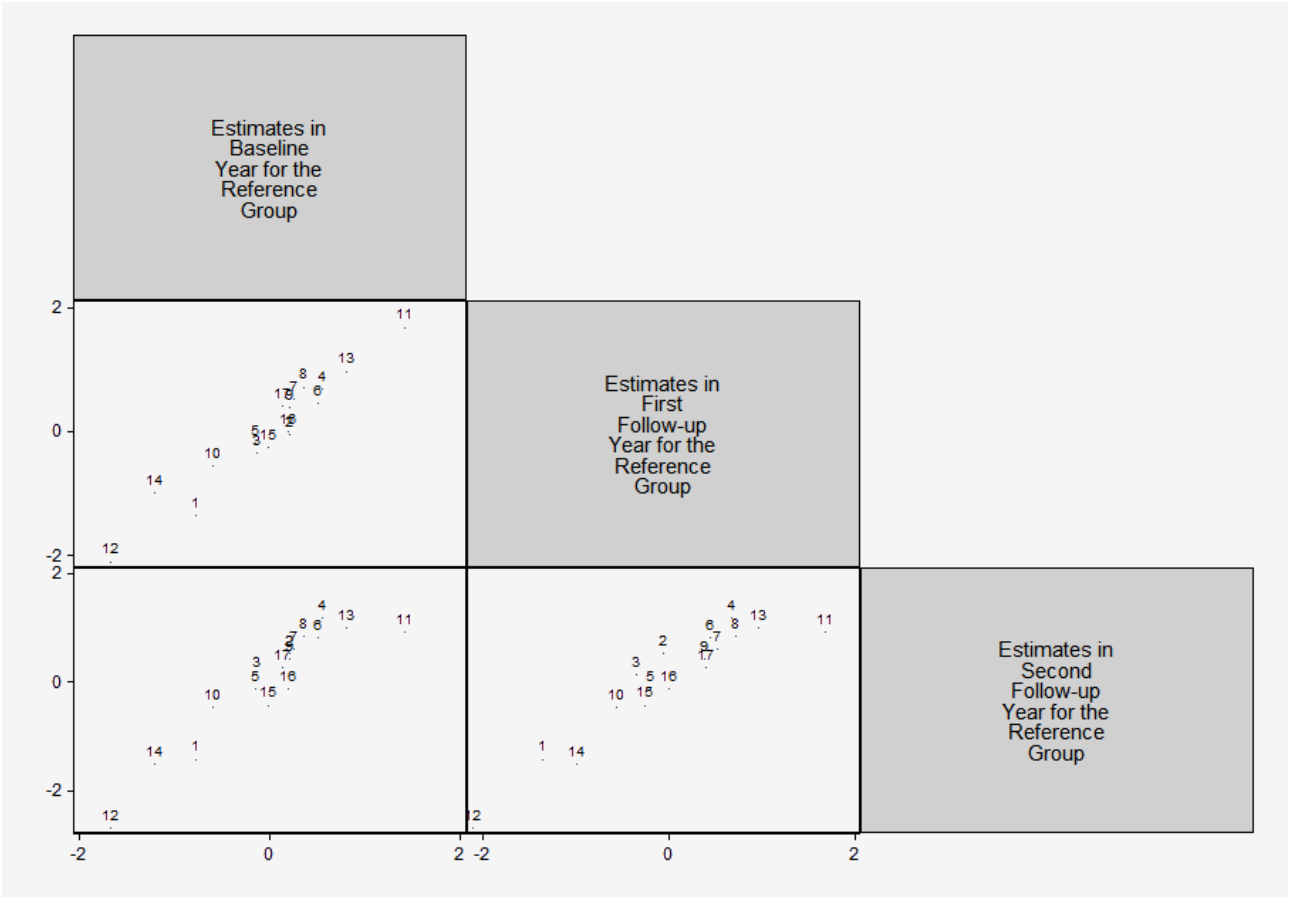


Figure 12: Relationship between item estimates for each year in the reference group. r_{01} = 0.96, r_{02}= 0.91, r_{12}=0.93

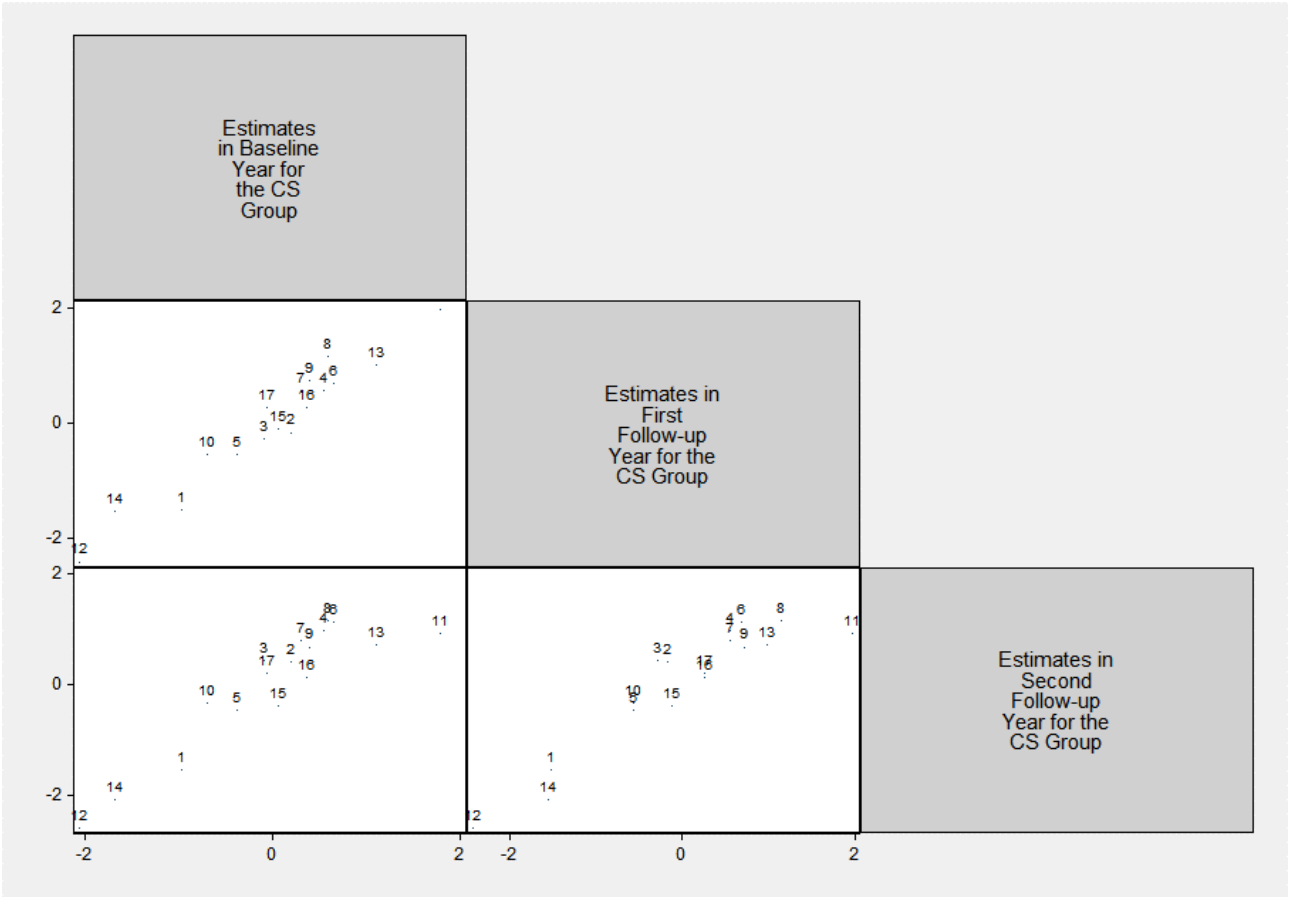


Figure 13: Relationship between item estimates for each year in the culturally sensitive group. ρ_{01}= 0.95, ρ_{02}= 0.90, ρ_{12}=0.84

Table 16: Item Parameter Estimate Differences for the Full Sample

Item	Interval 1	Interval 2	Overall
1	-0.55*	-0.05	-0.60*
2	-0.30	0.59*	0.29
3	-0.20	0.57*	0.37
4	0.07	0.48	0.56*
5	-0.07	0.11	0.04
6	-0.02	0.39	0.37
7	0.27	0.15	0.42
8	0.46	0.06	0.51
9	0.24	-0.01	0.24
10	0.08	0.13	0.21
11	0.22	-0.89**	-0.67*
12	-0.39	-0.50*	-0.89*
13	0.02	-0.07	-0.05
14	0.23	-0.51*	-0.28
15	-0.21	-0.22	-0.42
16	-0.16	-0.15	-0.31
17	0.31	-0.09	0.22

Item difference $|d| > 0.5$ logits*
Outliers of item difference distribution**

Table 17: Item Parameter Estimate Differences for Reference Group

Item	Interval 1	Interval 2	Overall
1	-0.57*	-0.08	-0.64*
2	-0.26	0.58**	0.32
3	-0.20	0.49	0.29
4	0.13	0.51**	0.64*
5	-0.03	0.07	0.04
6	-0.04	0.36	0.32
7	0.29	0.09	0.37
8	0.37	0.12	0.49
9	0.18	0.05	0.23
10	0.04	0.09	0.12
11	0.27	-0.75**	-0.48
12	-0.43	-0.59**	-1.02
13	0.17	0.04	0.21
14	0.23	-0.55**	-0.31
15	-0.23	-0.18	-0.40
16	-0.19	-0.13	-0.32
17	0.28	-0.13	0.14

Item difference $|d| > 0.5$ logits*
Outliers of item difference distribution**

Table 18: Item Parameter Estimate Differences for Culturally Sensitive Group

Item	Interval 1	Interval 2	Overall
1	-0.55*	-0.02	-0.57*
2	-0.36	0.57*	0.20
3	-0.19	0.70*	0.51*
4	0.01	0.41	0.42
5	-0.18	0.09	-0.09
6	0.03	0.43	0.46
7	0.25	0.23	0.49
8	0.56*	0.00	0.56*
9	0.33	-0.05	0.28
10	0.14	0.21	0.35
11	0.19	-1.04**	-0.85*
12	-0.36	-0.18	-0.54*
13	-0.11	-0.28	-0.38
14	0.15	-0.56*	-0.41
15	-0.17	-0.29	-0.45
16	-0.10	-0.14	-0.24
17	0.34	-0.07	0.27

Item difference $|d| > 0.5$ logits*
Outliers of item difference distribution**

The distributions of the differences are also shown as box plots in Figure 14 by interval and across the three samples. Outliers are also highlighted in the corresponding tables. The largest differences for the Full Sample were for Items 1, 4, 11, and 12. These differences show that all of the items become easier over time. These items also have large changes in difficulty for the Reference Group and Culturally Sensitive Group. These large shifts must be interpreted with caution because their values might be inflated. Since the three waves of data were estimated separately, they are not constrained to be on the same scale.

DIF model.

Another approach to quantify item difficulty parameter differences is to model them with a special parameter. For instance, if items are thought to work differently for different groups of students, a DIF parameter that identifies these student groups can be incorporated. In this case, the item difficulty differences are related to time-wise drift. Thus, a differential item function (DIF) parameter can be added to the item response model. The item-by-time DIF parameter is an interaction term between the item and the measurement occasion. The item difficulties represent the portion of item difficulty that remains constant over time. The item-by-year parameters show the portion of item difficulty that changes over time. In this formulation, the interpretation is straightforward. A larger item-by-year DIF parameter means that the item shifts more over time. Finally, since they are modeled simultaneously in one step, they are on the same scale and easily comparable to each other. As is will be described below, the item-by-year DIF parameters show clear evidence of item drift. Now that the difference parameters are all on the same scale, interesting and different patterns arise between the measurement occasions and analysis samples.

The item-by-year DIF parameters for the Full Sample are shown in Table 19. For this sample, the DIF parameters are smallest during the first follow-up. There appears to be approximately equal numbers of items with large DIF parameters in the baseline and second follow-up years. In this case, the interpretation could be that the items reach their mean difficulty in the middle measurement occasion, or when students are in Grade 10. In any previous year, most of the DIF parameters are positive, and some largely so. This could mean that in Grade 8 these items are actually more difficult. Conversely, the DIF parameters for the second follow-up are all negative, and the majority of them are large. This could mean the opposite: that the *constant* items are actually easier to students when they reach third round of data collection (Grade 12). There are multiple explanations for this phenomenon including student ability to remember items from previous waves, student familiarity with the topic, and curriculum timing. It is important to note that the items do not appear to have equal difficulty across the measurement occasions; the longitudinal measurement variance assumption is violated. Additionally, different patterns in the size of DIF parameters are seen for the Reference and Culturally Sensitive groups that do not follow the pattern of the full sample.

The DIF parameter estimates for the Reference and Culturally Sensitive groups are shown in Tables 20 and 21. Again, differences over 0.5 logits on the History scale are emphasized in the tables. For the Reference Group, this cut-off of 0.5 logits meant that

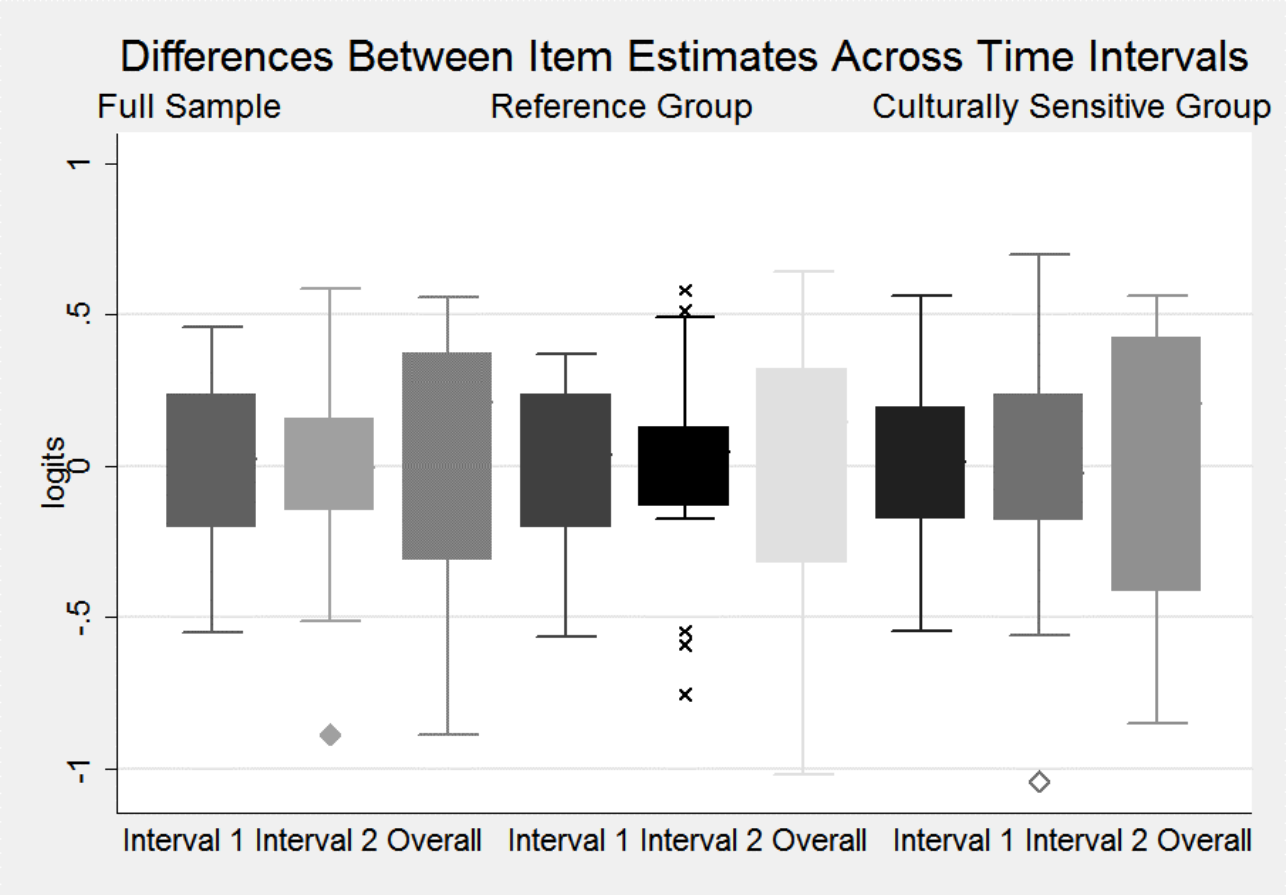


Figure 14: Differences in item parameter estimates across years in each sample group

Table 19: DIF Parameters for Full Sample

Item	Item**BY*	Item**F1*	Item**F2*
1	0.90*	-0.10	-0.80*
2	0.56*	-0.22	-0.34
3	0.49	-0.18	-0.31
4	0.36	-0.06	-0.30
5	0.55	0.01	-0.56*
6	0.45	-0.06	-0.39
7	0.33	0.12	-0.44
8	0.24	0.21	-0.45
9	0.40	0.16	-0.55*
10	0.44	0.06	-0.49
11	0.74*	0.44	-1.17*
12	0.93*	0.10	-1.03*
13	0.58*	0.11	-0.69*
14	0.53*	0.32	-0.84*
15	0.75*	0.07	-0.82*
16	0.71*	0.06	-0.77*
17	0.37	0.21	-0.58*

Item difference |dl| > 0.5 logits*

Table 20: DIF Parameters for Reference Group

Item	Item**BY*	Item**F1*	Item**F2*
1	0.36	-0.20	-0.16
2	0.01	-0.27	0.26
3	-0.01	-0.22	0.24
4	-0.23	-0.11	0.33
5	0.01	-0.03	0.02
6	-0.06	-0.12	0.18
7	-0.19	0.08	0.11
8	-0.26	0.10	0.16
9	-0.11	0.06	0.05
10	-0.06	-0.02	0.08
11	0.09	0.37**	-0.46
12	0.30	-0.07	-0.23
13	-0.10	0.07	0.03
14	-0.04	0.23	-0.19
15	0.22	-0.02	-0.20
16	0.19	-0.02	-0.17
17	-0.12	0.15	-0.03

Item difference |dl| > 0.5 logits*

Table 21: DIF Parameters for Culturally Sensitive Group

Item	Item*_BY_	Item*_F1_	Item*_F2_
1	0.24	-0.26	0.02
2	0.12	-0.29	0.17
3	-0.05	-0.27	0.32
4	-0.05	-0.09	0.15
5	0.08	-0.10	0.01
6	-0.07	-0.09	0.15
7	-0.17	0.05	0.12
8	-0.28	0.24	0.04
9	-0.13	0.17	-0.04
10	-0.18	-0.01	0.19
11	0.34	0.45*	-0.79*
12	-0.15	-0.23	0.38
13	0.27	0.09	-0.35*
14	-0.20	0.12	0.08
15	0.23	0.03	-0.26
16	0.17	0.03	-0.20
17	-0.16	0.16	0.00

Item difference $|d| > 0.5$ logits

only one of the DIF parameters, Item 11, was large in the first follow-up year. In fact it was large and positive, meaning that Item 11 was more difficult in that case. All other DIF parameters were close to zero. For the Culturally Sensitive Group, three item instances were identified as having large DIF parameters overall. Item 11 was identified again as having a large DIF parameter in both the first and second follow-ups. It is important to note that the sign of the DIF parameter changes. In the first follow-up the value is positive. In the second follow-up the value is negative. This pattern represents an item shift occurring in both directions. This type of shift, when compared with a uni-directional shift, could represent an even more important violation of longitudinal measurement invariance because values cannot just be scaled up or down with each additional wave of data collection. A summary of the time-wise DIF parameters is shown in Figure 15. The patterns of the shifts are discussed further in the next section. Given that there are shifts in the item difficulty, it is important to explore the nature of the patterns to gain insight into why they exist, why patterns between groups might be different, and to explore if longitudinal models still might be fit.

Patterns in item shift.

With three waves of data, there are several possible patterns of item difficulty shifts that could arise. Using the results from the separate calibrations and the DIF models, several possible patterns present themselves. Certain items become easier after each time interval. In the results from the separate calibrations, these items have negative difference values in Interval 1, meaning that the item difficulty decreased over interval one, and the same for Interval 2, meaning that the item difficulty also decreased over Interval 2. It follows that the overall difference is also negative. The other source of evidence, the item-by-year DIF parameters, can tell a similar story. To mirror a situation of decreasing difficulty, the item-by-year DIF parameters should result in decreasing difficulty after each interval when added to the base item parameter. When looking at Tables 16, 17, and 18, it is clear that across the samples, the same items exhibit this pattern (Items 1, 15, 16). This pattern might occur when students increase their understanding of certain topics as they progress through school, possibly because they learn a certain subject that they had not previously been taught. Two of the items that exhibit this pattern across all three of the samples involve information about the U.S. Supreme Court (Items 15 and 16). The item-by-year DIF parameters for all three of the groups shown in Tables 19, 20, and 21 also show that the difficulty for these same items decreases over time (Items 15 and 16). This material is generally covered in government or American law courses near the end of high school. It makes sense that these items would become easier when students take this course. It may seem feasible that the items would become easier, but the model assumes that they remain constant as a means of estimating student growth.

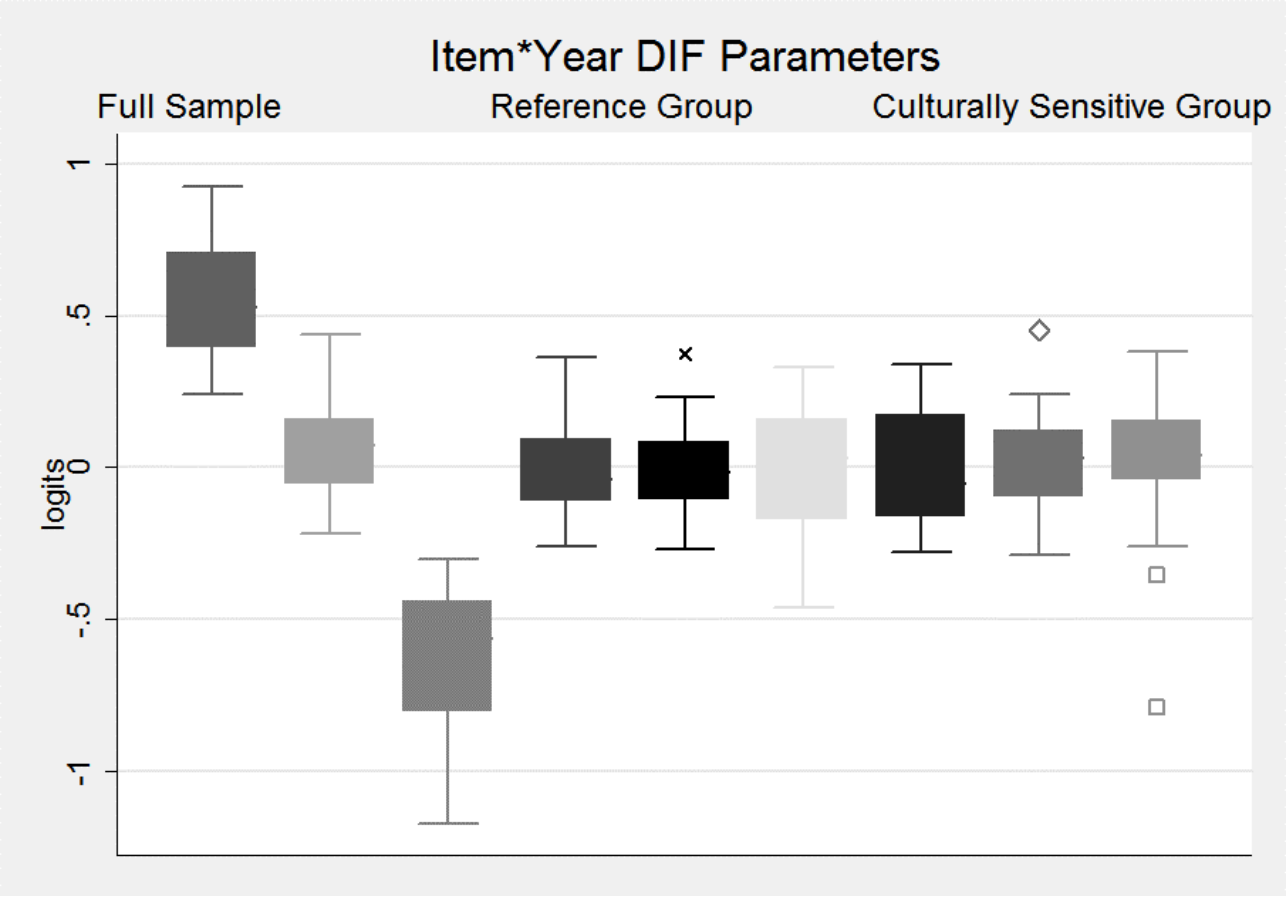


Figure 15: Item-by-year DIF parameters by group. Item*Year = Item and year interaction term

The same pattern of decreasing difficulty is seen for the Full Sample and the Reference Group for an item about a defendant's rights in the American judicial system (Item 12). However, for the Culturally Sensitive Group, this item shows a less clear pattern. The Culturally Sensitive Group might be becoming increasingly aware of discriminatory practices in the judicial system, as students mature into adulthood by their senior year of high school. The evidence from differences in item estimates from the separate calibrations shows that some items become less difficult. Plausible explanations for these item shifts can be drawn. Some researchers consider group effects without also considering item difficulty shifts when estimating growth. A more nuanced description of the assumption violation, which could be practically important in some growth modeling studies, can be presented if the combination of item shifts and differences in the patterns across groups is considered.

Just as some items become easier within certain groups, some items become more difficult over time. When examining results from the separate calibrations (Tables 16, 17, and 18), four items exhibit this pattern for all three samples (Items 4, 7, 8 and 10). Each difference estimate from the separate estimations is positive. The item-by-year DIF parameters would show a series of increasing values when added to the main item parameter. This occurrence is less intuitive than the previous because it is generally accepted that students become smarter over time. If students forget material, however, this pattern might occur. One item asks about westward immigration to the United States (Item 4). This pattern is also illustrated by the item-by-year parameters for the Reference Group (Item 4, Table 20). Since this material is generally covered in junior high school, or in the early years of high school, students could forget as time passes. Also, if understanding of a topic becomes more complex, the multiple-choice distracters may also become more tempting. For two of these items, the distracters are meant to confuse students about the House and Senate. Distracters that apply to the Senate are given for the question about the House, whereas distracters that apply to the House are given for the questions about the Senate (Items 7 and 8). When students become increasingly aware of the branches of government, it could become more difficult to select the correct answer from a set of distracters. For the Reference Group, two additional items become more difficult over time (Items 9 and 13) as shown in Table 17. This evidence is also mirrored by the item-by-year DIF parameters for the Reference Group (Item 9, Table 20). One item deals with the true motivation behind the voter registration act. The multiple-choice distracters cover political, racial, intellectual, and economic concerns. The correct answer involves discrimination. Since the Reference Group score lowered on the culturally sensitive items, this item could become more difficult for them over time, whereas this issue may remain salient to the Culturally Sensitive Group over time. The evidence from both the separate calibrations and the item-by-year DIF model shows that some items become more difficult for certain groups over time. Some would argue that this type of one-dimensional change could be accounted for within the growth model. However, for the example data from NELS, consisting of only three waves of collection, items actually shift in both directions. This type of change would require even more adjustments to be made to fit the typical growth model.

Some items become less difficult after the baseline year and then again more difficult after the first follow-up. This might occur when material is covered during the test administration time span. For instance, if a certain topic is covered in Grade 9, it may be easier during the first follow-up (Grade 10). If students forget this material by the second follow-up (Grade 12) it may become again more difficult. Three items show this pattern across the three samples as shown by

the separate calibration estimate differences (Items 2, 3, and 5). Two of these items are about the early peoples of the United States. These topics may be covered in introductory American history courses in Grade 9 or 10 and, unfortunately, discussed little after then. This would explain why the items first become easier and then more difficult. Conversely, other items become more difficult after the baseline year and less difficult after the first follow-up. Two items exhibit this pattern, as shown in by the difference estimates from the separate calibrations (Items 11 and 14). Evidence from the item-by-year DIF parameters is less obvious because it requires closer examination of the resulting item parameters in each year.

Authors differ on the technique to resolve the problem of items that drift in difficulty over time. One strategy is to eliminate them for a certain population or time period and use the remaining indicators (Scientific Software International, 2002). Another is to include as few as one common item between the groups that works well for both populations and time periods (Reise et al., 1993). However, in many practical settings, the issue of non-comparable common items is simply overlooked, placing too much faith in the item content. Individual items can provide important information even though they cannot be assumed to be longitudinally invariant. This issue is especially important in the case of common items that are responsible for fixing the scale across groups or time points. With more attention given to this topic from similar studies, techniques for resolving this issue will surface, thus strengthening the growth models used for analysis of psychological and educational constructs. Next, differences in the LG-IRM model are examined, by first ignoring item shift and then by allowing the item parameters to be estimated freely across years.

LG-IRM parameter estimates

Previous simulation studies showed that the item parameters and person parameters are well recovered by the LG-IRM model when item difficulties remain constant over time (Wilson, Zheng, & Walker, 2007). Parameter recovery estimates for the items were above 0.999 with a simulated sample of 100 common items. Parameter recovery estimates for the persons were above 0.989. Under the LG-IRM, one set of item parameters is estimated for all measurement occasions. The model is a within item multidimensional model where all items fall on a baseline dimension and items from the follow-up administrations also fall on the growth dimension. In addition, the variance components for the relationship between the baseline and growth dimension are also estimated. However, when item difficulties are known to vary over time, the results from the LG-IRM model should be examined carefully. If differences in the item parameters over time are not accounted for over time, the standard LG-IRM model could produce biased estimates.

Although the results from the basic approaches show that items do change over time, the LG-IRM model is estimated first ignoring these differences for comparison with a model that does incorporate these differences. These results are shown in Table 22. The variance components for the relationship between the baseline and growth dimensions are also shown. The correlation between the baseline and growth dimension, $Corr_{BG}$ is small and positive in the full and Culturally Sensitive Group. This implies that within these samples, those who do well on the base year test tend to do better over time. However for the Reference Group, the correlation between the baseline and growth dimension is negative and almost zero. This implies that, for those in the group that score low on the culturally sensitive items, those who score well on the base year administration do more poorly over time. These results should be interpreted cautiously because the item parameters have been constrained to be the same over time.

Next, the item parameters were estimated freely for each time point, still using the LG-IRM formulation: We will call this the *extended* LG-IRM. This approach was taken by te Marvelde, Glas, Van Landeghem, and Van Damme (2006) in their demonstration of the use of multidimensional models for longitudinal growth modeling**.** The estimates from this estimation are shown in Table 23 for the Full Sample, Table 24 for the Reference Group, and Table 25 for the Culturally Sensitive Group. Again, differences between item estimates larger than 0.5 logits were identified. For the Full Sample, three items show large differences in the second measurement occasion (Items 11, 12, and 13). For the Reference Group an overlapping group of three items shows a shift of 0.5 logits or more (Items 6, 11, and 13). For the Culturally Sensitive Group, a still overlapping set of four items showed big differences (Items 10, 11, 12, and 14). These items were also identified in the separate analyses shown above. This gives evidence that the extended model can model growth and also capture item shifts. Since it can capture item shifts, violations of longitudinal measurement invariance can be accounted for with the growth model. In accounting for item differences, the model produces more robust growth estimates as well as yearly information about the item. In comparison to traditional models, the LG-IRM produces more information and better growth estimates.

Table 22: Estimates from the LG-IRM Models

Item	Full	Reference	Culturally Sensitive
1	-1.09	-1.07	-1.09
2	0.22	0.25	0.16
3	-0.04	-0.08	0.02
4	0.77	0.83	0.68
5	-0.22	-0.12	-0.38
6	0.71	0.62	0.81
7	0.52	0.49	0.54
8	0.80	0.66	0.95
9	0.48	0.38	0.60
10	-0.51	-0.50	-0.49
11	1.48	1.32	1.62
12	-1.96	-1.96	-1.90
13	0.97	0.95	0.98
14	-1.27	-1.15	-1.48
15	-0.12	-0.17	-0.04
16	0.17	0.07	0.31
17	-0.58	-0.42	-0.77
Variance Components			
var(B)	0.90	0.47	0.78
var(G)	0.26	0.24	0.20
$cov(B,G)$	0.11	-0.01	0.11
ρ_{BG}	0.23	-0.03	0.27

In the extended version of the LG-IRM model, the variance components shown in Tables 23, 24, and 25 are comparable to those of the constrained model (Table 22) except for the variance of the growth dimension. Across all three sample groups the variance of the growth dimension is smaller when the item parameters are freely estimated. Furthermore, the differences in the variance components between the freely estimated items and the constrained model for the Culturally Sensitive Group are larger. This could mean the growth estimates are exaggerated if violations of longitudinal measurement invariance are ignored and that they could be exaggerated even more for non-reference groups. This evidence suggests that the extended LG-IRM can produce more precise measures of student growth across time for all groups of students. Since growth estimates serve as a measure of student learning over time, it is important to ensure that they are accurate. The results from the extended LG-IRM suggest that growth estimates can be made more accurate by incorporating item shifts versus ignoring them, which provides better measures of student learning over time.

Table 23: Extended LG-IRM Full Sample Estimates

Item	BY	F1	F2
1	-1.24	-1.13	-1.02
2	-0.19	-0.49	-0.39
3	-0.50	-0.60	-0.46
4	0.16	-0.13	-0.18
5	-0.64	-0.60	-0.62
6	0.19	-0.16	-0.23
7	-0.11	-0.17	-0.32
8	0.08	0.02	-0.22
9	-0.09	-0.17	-0.37
10	-1.02	-0.71	-0.69
11	1.23	0.48*	-0.24
12	-2.18	-1.52*	-1.43
13	0.57	0.05*	-0.25
14	-1.75	-1.00	-1.09
15	-0.37	-0.53	-0.68
16	-0.11	-0.38	-0.56
17	-0.34	-0.26	-0.46
Variance Components			
var(B)	0.88		
var(G)	0.14		
cov(B,G)	0.12		
corr(B,G)	0.33		

Note. BY = Base year. F1= first follow-up. F2 = second follow-up.

*Item difference $|d| > 0.2$ logits.

Table 24: Extended LG-IRM Reference Group Estimates

Item	BY	F1	F2
1	-0.66	-0.78	-0.75
2	0.32	-0.14	-0.10
3	-0.03	-0.29	-0.23
4	0.67	0.22	0.12
5	-0.04	-0.21	-0.32
6	0.62	0.11*	-0.01
7	0.36	0.15	-0.07
8	0.47	0.24	0.00
9	0.33	0.08	-0.13
10	-0.47	-0.39	-0.43
11	1.53	0.72*	0.03*
12	-1.55	-1.15	-1.16
13	0.92	0.37*	0.06
14	-1.09	-0.60	-0.78
15	0.10	-0.24	-0.42
16	0.31	-0.11	-0.32
17	0.25	0.09	-0.19

Variance Components	
var(B)	0.44
var(G)	0.12
$cov(B,G)$	0.02
$corr(B,G)$	0.09

Note. BY = Base year. F1= first follow-up. F2 = second follow-up.

*Item difference $|d| > 0.2$ logits.

Table 25: Extended LG-IRM Culturally Sensitive Group Estimates

Item	BY	F1	F2
1	-1.88	-1.53	-1.33
2	-0.72	-0.86	-0.69
3	-0.99	-0.91	-0.68
4	-0.36	-0.50	-0.50
5	-1.29	-1.05	-0.98
6	-0.25	-0.43	-0.45
7	-0.60	-0.49	-0.56
8	-0.32	-0.20	-0.44
9	-0.51	-0.41	-0.60
10	-1.60	-1.05*	-0.94
11	0.88	0.21*	-0.51
12	-2.96	-1.97*	-1.69
13	0.19	-0.28	-0.59
14	-2.59	-1.54*	-1.52
15	-0.85	-0.83	-0.95
16	-0.54	-0.64	-0.78
17	-0.98	-0.64	-0.76
Variance Components			
var(*B*)	0.89		
var(*G*)	0.12		
cov(*B*,*G*)	0.08		
corr(*B*,*G*)	0.26		

Note. BY = Base year. F1= first follow-up. F2 = second follow-up.

*Item difference $|d| > 0.2$ logits.

Discussion

This chapter discussed the prevalence of violations of longitudinal measurement invariance in studies where growth models are employed. Few studies within item response modeling circles consider the instability of items when they are being using in a multiyear study. Given the advantages of growth modeling, and more specifically, the possibilities of the LG-IRM, this paper discussed methods in which the LG-IRM could still be estimated, even with violation of longitudinal measurement invariance.

Example data from NELS were used to verify the existence of item shifts. Separate calibrations of three waves of data from NELS showed that items do indeed shift in difficulty over time. The DIF model parameter estimates also illustrated that items shift over time. Both types of analysis additionally showed that these items show different difficulty shifts across subpopulations. Since item shifts exist in such a widely circulated and analyzed data set, item shifts should be considered carefully.

Given evidence of item shift, it was assumed that there would be differences between the estimates from the LG-IRM and the *extended* LG-IRM. Through a comparison of the results from these two models, findings revealed that growth estimates were exaggerated when violations of longitudinal measurement invariance are ignored. In addition, estimates from the LG-IRM were most different for the non-reference group. These results suggested that if violations of longitudinal measurement invariance are found, then the *extended* LG-IRM can produce more robust estimates by incorporating item shifts versus ignoring them. The variance of the growth dimension was less exaggerated, which made the growth estimates more precise. These more robust and more precise growth estimates provided better measures of true student growth over time.

This discussion is based on results from an example data set. Future work in this area could illustrate the performance of this model under different simulated events. Such a simulation study would present the opportunity to describe the effects of such violations under many conditions. For instance, in educational testing, the set of common items may be smaller than the set of items given in any year. This study showed an example in which only common items were used. A simulated data set including some common and unique items could be studied. This data set could provide the chance to explore whether the negative effects of item shift can be balanced out with unique items.

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit model. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Thoery into practice* (pp. 143-166). Norwood, NJ: Ablex.

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

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Bachman, J. G., & Johnston, L. D. (1978). *The Monitoring the Future Project: Design and Procedures.* Ann Arbor: University of Michigan, Institute for Social Research.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-addedassessment of teachers. *Journal of Educational and Behavioral Statistics*, 29, 37-65.

Baumeister, R. (2005). *The Cultural Animal.* New York: Oxford University Press.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Blascovich, J. & Tomaka, J. (1991). Measures of self-esteem. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 161-194). New York: Academic Press.

Block, J., & Robins, R. W. (1993). A longitudinal study of consistency and change in self-esteem from early adolescence to early adulthood. *Child Development, 64*, 909–923.

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

Bock, R. D., & Zimowski, M. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer-Verlag.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12,* 261-280.

Bond, T. G., Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

Briggs, D. C., & Weeks, J. P. (2008, March). *The Persistence of value-added school effects.* Paper presented at the 2008 Annual Meeting of the American Educational Research Association, New York, NY.

Briggs, D. C., Weeks, J. P., & Wiley, E. W. (2008, April). *The sensitivity of value-added modeling to vertical scaling.* Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87-100.

Byrne, B., Shavelson, R., & Muthen, B. (1989). Testing for partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Cagnone, S., Moustaki, I., & Vasdekis, V. (2009) Latent variable models for multivariate longitudinal ordinal responses. *British journal of mathematical and statistical psychology*, *62,* 401-415.
Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313*, 1307-1310.
Conroy, D. E., Metzler, J. N., & Hofer, S. M. (2003). Factorial invariance and latent mean stability of performance failure appraisals. *Structural Equation Modeling, 10*, 401–422.
Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology, 85*, 507-516.
Crocker, J., Luhtanen, R. K., Cooper, M. L., & Bouvrette, S. (2003). Contingencies of self-worth in college students: Theory and measurement. *Journal of Personality and Social Psychology, 85*, 894-908.
Crocker, J., & Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin, 130*, 392–414.
Crocker, J., Sommers, S., & Luhtanen, R. (2002). Hopes dashed and dreams fulfilled: Contingencies of self-worth in the graduate school admissions process. *Personality and Social Psychology Bulletin*, 28, 1275-1286.
Crocker, J., & Wolfe, C. T. (2001). Contingencies of self-worth. *Psychological Review*, 108, 593-623.
Diggle, P. J., Heagerty, P., Liang, K.Y., Zeger, S. L. (2002). *Analysis of longitudinal data.* Oxford, UK: Oxford University Press.
Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*, 227-246.
Embretson, S. (1991). A multidimensional latent trait model for measuring learning change. *Psychometrika, 56*, 494-515.
Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Gray-Little, B., & Hafdahl, A. R. (2000). Factors influencing racial comparisons of self-esteem: Aquantitative review. *Psychological Bulletin, 126*, 26-54.
Gray-Little, B., Williams, V.S.L., & Hancock, T. D. (1997). An Item Response Theory Analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin. 23*, 443-451.
Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for item responsetheory itemparameters using separate versus concurrent estimation in the common-item equatingdesign. *Applied Psychological Measurement*, *26,* 3-24.
Harter, S. (1993). Causes and consequences of low self-esteem in children and adolescents. In R.F. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 87-116). New York: Plenum Press.
Holland, P. W., & Rubin, D. B. (Eds.). (1985). *Test equating*. New York, NY: Academic Press.
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.
Hughes, M., Demo, D. H. (1989). Self-Perceptions of Black Americans: Self-Esteem and Personal Efficacy. *American Journal of Sociology, 95,* 132-159.
Jackson, J.S., & Neighbors, H. W. (1996). *National Survey of Black Americans, Waves 1-4,*

1979-1980, 1987-1988, 1988-1989, 1992. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
Kawash, G.F., & Scherf, G.W. (1975). Self-esteem, locus of control, and approval motivation in married couples. *Journal of Clinical Psychology, 31,* 715-720.
Kolen, M. J., & Brennan, R. L. (1995). *Test equating, scaling, and linking: Methods and practices* (1st ed.). New York: Springer-Verlag.
Lambert, M., Essau, C., Schmitt, N., & Samms-Vaughan, M. (2006). Dimensionality and psychometric invariance of the Youth Self-report Form of the Child Behavior Checklist in cross-national settings, *Assessment*, *14,* 231-245.
Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.
Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47-68.
Long, J. D., Harring, J. R., Brekke, J. S., Test, M. A., & Greenberg, J. (2007). Longitudinal construct validity of Brief Symptom Inventory subscales in schizophrenia. *Psychological Assessment, 19,* 298-308.
Lord, F. M. (1981). *Standard error of equating by item response theory*. Princeton, NJ: Educational Testing Service: (Tech. Rep. No. RR-81–49).
Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaledstudent achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35-62.
Masters, G. (1981). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
McCarthy, J. D., & Hoge, D.R. (1982). Analysis of age effects in longitudinal studies of adolescent self-esteem. *Developmental Psychology, 18*, 372-379.
Meijer, R. (2002, January). Personality/Psychopathology Measurement and IRT: promising opportunities. Presented at Tilburg Univeristy.
Miller, J., Hoffer, T., Suchner, R., Brown, K., & Nelson, C. (1992). *LSAY codebook: Student, parent, and teacher data for cohort two for longitudinal years one through four (1987–1991).* DeKalb: Northern Illinois University.
Motl, R., Dishman, R., Birnbaum, A., & Lytle, L. (2005). Longitudinal invariance of the Center for Epidemiological Studies Depression Scale among girls and boys in middle school. *Education and Psychological Measurement, 65*, 90-108.
National Education Longitudinal Study of 1988: *Base-Year to Third Follow-up Data Files*. (2000). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
National Research Council and National Academy of Education. (2010). *Getting value out of value-added: report of a workshop.* Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, H. Braun, N. Chudowsky, & J. Koenig (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6,* 328-362.
No Child Left Behind Act of 2001, Pub L. No. 107-110. 115 Stat. 1425 (2002).
Phinney, J.S., & Chavira, V. (1992). Ethnic identity and self-esteem: An exploratory

longitudinal study. *Journal of Adolescence, 15*, 271–281.
Prasad, M.S., & Thakur, G.P. (1977). *Manual and directions for Self-esteem Inventory*. Agra: Agra Psychological Research Cell.
Pentz, A., & Chou, C. (1999). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting in Clinical Psychology, 62*, 450–462.
Rahu, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529.
Ralph, J., & Crouse, J. (1997). Reading and mathematics achievement: Growth in high school. Retrieved from http://www.ed.gov/NCES/pubs/.
Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008: Trends in academic progress. NCES 2009-479.* Jessup, MD: Ed Pubs.
Reise, S., Widman, K., & Pugh, R. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.
Rapkin, B.D., & Hirsch, B.J. (1987). The transition to junior high school: a longitudinal study of self-esteem, psychological symptomatology, school life, and social support. *Child Development, 58,* 1235-43.
Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Robins, R. W., Trzesniewski, K. H., Tracy, J. L., Gosling, S. D., & Potter, J. (2002). Global self esteem across the life span. *Psychology & Aging, 17*, 423-434.
Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton
Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement*, 14, 271-282.
Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75-92.
Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of the differencescore in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.
Rubin, D.B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics, 29*, 185–203.
Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging, 13*, 8–20.
Schmidt, W.H., Houang, R.T., & McKnight, C.C. (2005). Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 145-164). Maple Grove, MN: JAM Press.
Scientific Software International, Inc. (2002). *IRT from SSI: Bilog-MG, Multilog, Parscale, and Testfact.* Lincolnwood, IL: Scientific Software International, Inc.
Seltzer, M., Choi, K., & Thum, Y. M. (April, 2002). *Examining relationships between where students start and how rapidly they progress: Implications for constructing indicators that help illuminate the distribution of achievement within schools. (CSE Technical Report 560).* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/CRESST/Reports/TR560.pdf.
Simmons, R.G. (1978). Blacks and High Self-Esteem: A Puzzle. *Social Psychology, 41,*

54-57.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.
Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1,* 81 - 97.
Steele C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.
Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.
Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.
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*, 5-34.
Thum, Y. (2002). *Measuring student and school progress with the California API. CSE Technical Report No. 578.* University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Twenge, J. M., & Crocker, J. (2002). Race and self-esteem: Meta-analyses comparing Whites, Blacks, Hispanics, Asians, and American Indians and comment on Gray-Little and Hafdahl (2000). *Psychological Bulletin*, *128*, 371-408.
Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.
von Davier, M. (1995). *WINMIRA V 1.741 for Windows 3.x.* Kiel: IPN.
Wang, W. (1999). Direct estimation of correlations between latent traits within the IRT framework. *Methods of Psychological Research Online, 4.*
Wang, W. C., Wilson, M. R., & Adams, R. (1998). Measuring individual difference in change with multidimensional Rasch models. *Journal of Outcome Measurement, 2*, 240-265.
Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.
Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West, (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.
Willms, J. D. (2008). *Seven key issues for assessing 'value added' in education.* Paper at workshop on value-added assessment. Washington, DC: National Research Council and the National Academy of Education.
Wilson, M. R., Zheng, X., & Walker, L. (2007, April). *Latent growth item response models*. Presented at the BEAR Center Seminar, Berkeley, CA.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: Mesa Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: Mesa Press.

Wu, M., Adams, R., & Wilson, M. R. (1998). *ConQuest user guide.* Hawthorn, Australia: ACER.

Zeigler-Hill, V. (2007). Contingent self-esteem and race: Implications for the Black self-esteem advantage. *Journal of Black Psychology*, *33*, 51-74.

Zimmerman, M.A., Copeland, L.A., Shope, J.T., Dielman, T.E. 1997. A Longitudinal Study of Self-Esteem: Implications for Adolescent Development. *Journal of Youth and Adolescence*, *26*, 117-142.

Appendices

Appendix A

1. /* Generation of Baseline ability and Linear Growth Parameters. Only these outputs, contained in the ".abl" file are used. The simulated data and item difficulties do not match the simulation and are not used. Here N is the sample size. */

```
reset;
generate !nitems=2200:2200, npersons=N, maxscore=1,
abilitydist=mvnormal(-0.5:1:0.3:1:0.3)
>> simDontUse.dat, simDontUse.itm, simUse.abl;
```

2. /* Generation of Item Difficulties. Only these outputs, contained in the ".itm" file are used. The simulated data and person parameters do not match the simulation and are not used. */

```
reset;
generate !nitems=165, npersons=N, maxscore=1, itemdist=uniform(-2.8:2.8),
abilitydist=normal(0:1)
>> simDontUse.dat, simUse.itm, simDontUse.abl;
```

3. Generate Responses using the item difficulties and ability parameters.

/* Baseline Ability parameters and growth parameters were crossed with item difficulty parameters in excel. Formula (2) was then applied to produce the probability of a correct response. Probabilities over .5 were rounded to a correct response while probabilities below were rounded to an incorrect response.*/

```
Pijk=exp((ablj+k*growj)-diffi))/[1+ exp((ablj+k*growj)-diffi))]
Rijk=round(Pijk, 1)
```

Appendix B

```
/* Estimates the LG-IRM with vertical scaling design with the same number of items in each year and the same common items across years*/
reset;
title LG-IRM with Common Items on Vertical Scale;
datafile simUsedata.dat;
format responses 6-45, / 6-45, / 6-45, / 6-45, / 6-45,/ 6-45;
import designmatrix<<VS.des;
score (0,1) (0,1) ( )   !items(1-40);
score (0,1) (0,1) (0,1) !items(41-80);
score (0,1) (0,1) (0,2) !items(81-120);
score (0,1) (0,1) (0,3) !items(121-160);
score (0,1) (0,1) (0,4) !items(161-200);
score (0,1) (0,1) (0,5) !items(201-240);
model item;
estimate;
```

Appendix C

1. Useful Person: I am a useful person to have around.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (0), Often True (1), Not Often True (2), Never True (3)

2. Person of Worth: I feel that I'm a person of worth.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (0), Often True (1), Not Often True (2), Never True (3)

3. Can't Do Anything Right: I feel that I can't do anything right.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (3), Often True (2), Not Often True (1), Never True (0)

4. Life Not Useful: I feel that my life is not very useful.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (3), Often True (2), Not Often True (1), Never True (0)

5. Not Proud: I feel I do not have much to be proud of.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (0), Often True (1), Not Often True (2), Never True (3)

6. Does Good Job: As a person I do a good job these days.
Original codes: Almost Always True (1), Often True (2), Not Often True (3), Never True (4)
Recodes: Almost Always True (0), Often True (1), Not Often True (2), Never True (3)

Appendix D

```
Title Wave 0 RATING SCALE
datafile NSBAWide.dat;
format id 1-5 responses 7-12;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + step;
estimate;
Title Wave 1 RATING SCALE
datafile NSBAWide.dat;
format id 1-5 responses 13-18;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + step;
estimate;
Title Wave 2 RATING SCALE
datafile NSBAWide.dat;
format id 1-5 responses 19-24;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + step;
estimate;
Title Wave 3 RATING SCALE
datafile NSBAWide.dat;
format id 1-5 responses 25-30;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + step;
estimate;

Title Wave 0 PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 7-12;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + item*step;
estimate;
Title Wave 1 PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 13-18;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + item*step;
estimate;

Title Wave 2 PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 19-24;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + item*step;
estimate;
```

```
Title Wave 3 PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 25-30;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5);
model item + item*step;
estimate;

Title Wave 0 LR-PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 7-12 wave 33;
model item + item*step;
regression wave;
estimate;
Title Wave 1 LR-PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 13-18 wave 33;
model item + item*step;
regression wave;
estimate;


Title Wave 2 LR-PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 19-24 wave 33;
model item + item*step;
regression wave;
estimate;
Title Wave 2 LR-PARTIAL CREDIT;
datafile NSBAWide.dat;
format id 1-5 responses 25-30 wave 33;
model item + item*step;
regression wave;
estimate;

Title LG-IRM Self-Esteem;
datafile NSBALong.dat;
format pid 1-5 responses 13-18/ 13-18/ 13-18/1 3-18;
recode (1, 2, 3, 4) (0, 1, 2, 3) !item(1,2,6, 7, 8, 12, 13, 14, 18, 19, 20,
24);
recode (1, 2, 3, 4) (3, 2, 1, 0) !item(3,4,5, 9, 10, 11, 15, 16, 17, 21, 22,
23);
import designmatrix <<LGIRMse.des;
score (0,1, 2, 3) (0,1,2,3) ()!)(items 1-6);
score (0,1, 2, 3) (0,1, 2, 3) (0, 1, 2, 3)! (items 7-12);
score (0,1, 2, 3) (0,1, 2, 3) (0, 2, 4, 6)! (items 13-18);
score (0,1, 2, 3) (0,1, 2, 3) (0, 6, 8, 10)! (items 19-24);
model item + item*step;
estimate;
```

Appendix E

```
* Creates Wright Maps using MLE, WLE, and EAP estimates
twoway (scatter difficulty item, ytitle(logits) jitter(.5) msymbol(circle)
mcolor(black) legend(off)) (histogram mle_t if mle_t>=-2&mle_t<=4,
xtitle(Frequency (MLE)) freq horizontal fcolor(none) by(year) lcolor(gs9)
legend(off))
twoway (scatter difficulty item, ytitle(logits) jitter(.5) msymbol(circle)
mcolor(black) legend(off)) (histogram wle_t if wle_t>=-2&wle_t<=4,
xtitle(Frequency (WLE)) freq horizontal fcolor(none) by(year) lcolor(gs9)
legend(off))
twoway (scatter difficulty item, ytitle(logits) jitter(.5) msymbol(circle)
mcolor(black) legend(off)) (histogram eap_t if eap_t>=-2&eap_t<=4,
xtitle(Frequency (EAP)) freq horizontal fcolor(none) by(year) lcolor(gs9)
legend(off))
*Creates plot of individual trajectories
twoway (line eap_t year if person<40, ytitle(Low Self-Esteem (logits)))
```