An IRT Model to Estimate Differential Latent Change Trajectories in a Multi-Stage, Longitudinal Assessment

A Thesis
Presented to
The Academic Faculty

By

Hi Shin Shim

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in the
School of Psychology

Georgia Institute of Technology

May 2009

An IRT Model to Estimate Differential Latent Change Trajectories in a Multi-Stage, Longitudinal Assessment

Approved by:

Dr. James S. Roberts, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Susan Embretson
School of Psychology
Georgia Institute of Technology

Dr. Larry James
School of Psychology
Georgia Institute of Technology

Date Approved: March 23, 2009

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Repeated measures designs are widely used in educational and psychological research to compare the changes exhibited in response to a treatment. Traditionally, measures of change are found by calculating difference scores (subtracting the observed initial score from the final score) for each person. However, problems such as the reliability paradox and the meaning of change scores arise from using simple difference scores to study change. A new item response theory model will be presented that estimates latent change scores instead of difference scores, addresses some of the limitations of using difference scores, and provides a direct comparison of the mean latent changes exhibited by different groups (e.g. females versus males). A simulation-based test was conducted to ascertain the viability of the model and results indicate that parameters of the newly developed model can be estimated accurately. Two sets of analyses were performed on the Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K) to examine differential growth in math ability between 1) male and female students and 2) Caucasian and African American students from kindergarten through fifth grade.

CHAPTER 1

INTRODUCTION

Repeated measures designs are widely used in educational and psychological research, where the same subjects/individuals are repeatedly measured over time on one or more observed variables. The primary purpose of using repeated measures designs is to compare the changes over time exhibited in response to a treatment. For example, in educational research, this methodology is important in assessing the amount of learning (i.e. change) that results from teaching (i.e. treatment) over time (e.g. year, semester). Traditionally, measures of change are found by calculating difference scores (subtracting the observed initial score from the final score) for each person. However, problems such as the reliability paradox and meaning of difference scores (Lord, 1956 & 1958) arise from using the simple difference of successive test scores to study change. For example, the reliability of a difference score is near zero when reliabilities of pretest and posttest are both high and when a large pretest-posttest correlation exists (Williams & Zimmerman, 1996). In addition, score variance, score correlations, and reliabilities are population dependent. Alternative methods such as residual change scores and multi-wave methods (Dimitrov & Rumril, 2003; Willett, 1989b; Rogosa & Willett, 1985) have developed in response to these problems, but the item response models (IRT) for the analysis of repeated measures designs appear the most promising in circumventing some of the classical problems in the measurement of change (Gluck & Spiel, 1997).

There are three common univariate IRT approaches in the analysis of repeated measures: 1) separate calibration, 2) concurrent calibration and 3) fixed parameter calibration techniques. However, unidimensional IRT models ignore the correlations between latent trait scores over time, yielding less precise estimates of latent trait scores (Roberts & Ma, 2006).

Multidimensional IRT models, on the other hand, account for the correlation in latent trait scores. Andersen (1985), Embretson (1991), Roberts & Ma (2006) and others have developed

multidimensional IRT (MIRT) models for the analysis of repeated measures. A new MIRT model will be presented in this thesis that allows for the estimation of the latent change scores and provides a direct comparison of the mean latent changes exhibited by different groups (e.g., females versus males).

The purpose of this study is to develop a new multidimensional item response theory (MIRT) model for repeated measurement analysis and to apply the new model to data from the Early Childhood Longitudinal Study - Kindergarten Cohort (ECLS-K; Rock & Pollack, 2002). The new model is a three-parameter logistic model for longitudinal assessment, in which multiple group structure is also incorporated into the model (e.g., Caucasian versus Asians over time). This model can also be constrained to yield a two-parameter logistic model for those items where it is unnecessary to estimate the pseudo-guessing parameter and in cases where mixed format tests are used (e.g. both multiple choice and free response questions in one test).

An initial, small-scale simulation test was conducted to ascertain the viability of the parameter estimation with the model. Both true item parameters and the simulation design were based on item responses from a real testing program, namely, the Early Childhood Longitudinal Study - Kindergarten cohort (ECLS-K; Rock & Pollack, 2002). Responses from 2000 simulees to a two-stage adaptive assessment were generated for six time points. Parameters were then estimated using a joint Bayesian estimation technique implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Dunn, 2007). Test results indicate the procedure can estimate the model parameters accurately.

The Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K) is an ongoing longitudinal study, where students are tested from kindergarten through eighth grade in the areas of science, mathematics, reading, and general knowledge. In each round of testing, there are two stages: the routing stage and second adaptive stage. In this thesis, the ECLS-K data are analyzed using the new model to determine if differential growth exists between 1) male and female students and 2) Caucasian and African American students in math ability. Item

parameters and differential growth estimates will be reported along with the item characteristic curves and test information curves derived from alternative test stages at different assessment times. Latent trait distributions and growth trajectories will be reported for each gender and each ethnicity.

CHAPTER 2

Distinguishing between Observed, True and Difference Scores in Classical Test Theory

In classical test theory, we attempt to discern the true nature of an individual's status from a fallible measure, the observed score. It is a linear combination of true score and measurement error. The measurement model is given as:

$$X_{nt} = \xi_{nt} + e_{nt} \quad (1)$$

Where n represents the nth individual, t is the time/occasion of test administration, X_{nt} is the observed score, ξ_{nt} is the true score, and e_{nt} is a random variable that constitutes measurement error. Note that e_{nt} is assumed to have a population mean equal to zero. If we are interested in measuring the change or growth, then parallel test forms are administered, for example, at two different time points to the same group of individuals. The observed change is simply the difference between the pretest and posttest score, written as:

$$D_n = X_{n2} - X_{n1} = (\xi_{n2} + e_{n2}) - (\xi_{n1} + e_{n1}) \quad (2)$$

$$D_n = (\xi_{n2} - \xi_{n1}) + (e_{n2} - e_{n1}) \quad (3)$$

$$D_n = \gamma_n + e_{\gamma_n} \quad (4)$$

Where D_n is the difference score for the nth individual, γ_n is the true change score, and $e_{\gamma_n} = e_{n1} - e_{n2}$ is the difference in error between the two test forms. It is important to differentiate between difference scores and true change scores. The true change is the difference between the true scores on the initial and final tests.

Three assumptions about the errors of measurement must be met in classical test theory: 1) the expected value of the error scores in the population is zero, 2) error scores are

uncorrelated with each other in the population, and 3) error scores are uncorrelated with true scores on either test in the population. Two assumptions, which are weakly met, must hold true for the use of parallel forms: 1) the true scores must be identical and 2) error variance must be the same for both tests (Lord, 1980). Difference scores can only be infallible measures of true change if and only if "the tests are perfectly reliable" (Lord, 1958). For example, when the measurement error (e_{γ_n}) is large and negative, the difference score can be negative even though the true change is positive (Lord, 1958).

CHAPTER 3

THREE MAJOR ISSUES WITH USING DIFFERENCE SCORES

3.1 Reliability Paradox

Difference scores are easy to calculate and an unbiased estimate of true change, but despite these properties, many methodologists have criticized the use of difference scores for several reasons, such as the reliability paradox. Reliability of a difference score is dependent on the correlation between the initial and final scores. The higher the correlation between the two test scores, the lower the reliability will be; all other things being equal. A lower correlation between test scores would increase the reliability of difference scores; however, we must then question whether the tests measure the same latent variable on the two different occasions (Lord, 1956; Bereiter, 1963). It would not be logical to compare the initial status to the final status of an individual and estimate the change if the tests are not measuring the same dimensions. This is true even when two identical tests are administered, where at the second administration the test no longer measures the same construct because the individual has changed so drastically (Lord, 1958). For example, a third grader initially fails to answer an item correctly due to the lack of skill. Then the same individual fails the same item in seventh grade because now, with the gain in knowledge and understanding of concepts, over-thinks or makes the problem harder than it is.

The reliability of a difference score is the ratio of true change variance and difference score variance for all people in the population and is calculated as:

$$\rho(D) = \frac{\sigma_{\gamma}^2}{\sigma_D^2} = \frac{\sigma_{\gamma}^2}{\sigma_{\gamma}^2 + \sigma_{e_{\gamma}}^2} \tag{5}$$

Where $\rho(D)$ is the reliability of the difference score, σ_{γ}^2 is the true change variance, $\sigma_{e_{\gamma}}^2$ is the error variance, and σ_D^2 is the difference score variance. From equation 5, note the connection between reliability and measurement error. If the random error variance is large relative to the inter-individual variation in true change score, the reliability is low. The reliability is high when random measurement error relatively smaller than the true change variation. Another way to examine reliability of difference scores is in terms of variances and reliabilities of constituent test scores, and correlations between initial and final observed scores:

$$\rho(D) = \frac{\sigma_{X_1}^2 \rho_{X_1} + \sigma_{X_2}^2 \rho_{X_2} - 2\sigma_{X_1}\sigma_{X_2}\rho_{X_1X_2}}{\sigma_{X_1}^2 + \sigma_{X_2}^2 - 2\sigma_{X_1}\sigma_{X\,2}\rho_{X_1X_2}} \tag{6}$$

Where $\sigma_{X_1}^2$ and $\sigma_{X_2}^2$ are variances of observed scores X_1 and X_2, respectively, ρ_{X_1} and ρ_{X_2} are the reliabilities of X_1 and X_2, respectively, and $\rho_{X_1X_2}$ is the correlation between the two observed scores over all people in the population. By examining equation 6, we find the two sources of the reliability paradox: reliabilities of observed scores and the correlation between the initial and final scores. The numerator and denominator look similar; however, the numerator will always be less than or equal to the denominator because the variance of each observed score is multiplied by its respective reliability. Consequently, higher values of ρ_{X_1} and ρ_{X_2} are desirable. Reliability of the difference score is also a function of the correlation between the initial and final scores such that, for a given level of constituent test reliability less than one, a higher correlation between test scores decreases the reliability of the difference score. However, researchers generally desire a high value for this correlation to ensure construct validity (Rogosa & Willett, 1985). Correlations between the initial and final observed scores will be high only when most of the individuals in the group are changing at approximately the same rate, maintaining the rank order (Rogosa & Willett, 1985; Willet, 1985). Under such conditions, the reliability of the

difference score is lowered due to the low variation of true change scores, making it is harder to distinguish between the individuals. In addition, if reliability of the difference score is low, then it is not appropriate to correlate the difference score with other variables in the population (Mellenbergh, 1999). Rogosa and Willett (1985) argue that different individuals change at different rates; therefore, their trajectories may crisscross and the rank ordering of the individuals will fluctuate naturally as time passes. This lowers the correlation but increases the reliability of difference scores, so they conclude that $\rho_{X_1X_2}$ should not be used as an index of construct validity but rather a measure of heterogeneity in change.

3.2 Correlations between the initial and change score

Another major criticism regarding difference scores is the negative correlation between the initial score and change score. Thorndike (1924) first noted the spurious negative correlation, also known as the Law of Initial Values (Rogosa & Willett, 1985), which implies individuals with low scores on the initial test will change faster or gain more than those with high initial scores. This may be true, but Lord also suggests an alternative reason for this occurrence: regression towards the mean. Given the effects of measurement error, for example, those with high initial status may have had a large measurement error in their score, but they were not as "lucky" on the second test administration, reducing their average score (Lord, 1958). Linn and Slinde (1977) pointed out that if the change score is not independent of initial status, then the measure of change should be considered unfair because it gives "an advantage to persons with certain values of the pretest scores." However, Willett (1989) disagrees and questions: "Why should change and status be unrelated?" He argues that we are a product of past change and that current changes influence our future status. Therefore, the correlation between our status and change is unavoidable.

3.3 Meaning of change scores

Only when initial and final measurements are perfectly reliable can difference scores provide infallible estimates of change. So how do we interpret difference scores when they are not? Sometimes we are interested in differences between individuals in addition to changes within an individual. Interpreting gain scores is relatively easy when comparing individuals with the same starting ability. However, a problem arises when attempting to compare individuals who start at different initial abilities. When a negative correlation between the initial status and the change score exists, high ability individuals gain less numerically than individuals of lesser ability (Lord, 1958). However, a small difference score for an individual with high initial status can actually represent greater true change than a comparable difference score for an individual with an average initial status (Lord, 1958; Embretson & Reise, 2000; Bereiter, 1963). If the distribution of initial test scores is normal or unimodal and intervention/treatment produces a negatively skewed distribution of posttest scores, then a ceiling effect occurs because the scores at the higher end are truncated (Williams & Zimmerman, 1996). With this compression at the upper end of the scale, even if these individuals change drastically, it would be "physically impossible for them to show any sizable gain on the post-test" (Lord, 1958). While people erroneously suggest that the scale is interval in nature, it cannot be assumed to possess numerically equal intervals unless empirically proven otherwise (Fischer, 2003; Lord, 1958), so lesser gains for high status individuals may actually represent greater gains than those made by individuals of moderate status (e.g. one point versus five points, respectively) (Embretson & Reise, 2000; Bereiter, 1963). Thus, gain scores have meaning only when comparing individuals who start at comparable initial status positions.

CHAPTER 4

ITEM REPONSE THEORY AS AN ALTERNATIVE TO THE USE OF DIFFERENCE SCORES

With these criticisms of difference scores, their usefulness has been and should be seriously questioned. Item response theory (IRT) models provide benefits and deal with some of the limitations of using difference scores when the models fit the data well. First, parallel tests, whose assumptions are rarely met, are not necessary (i.e. the true scores and standard deviations of the pretest and posttest do not have to be same). Embretson and Reise (2000) and Feldt and Brennan (1989) considered it more realistic to assume parallel tests for profile settings than in growth settings, where standard deviations of test scores are likely to vary due to treatment. Second, the precision of the IRT model parameters is calculated instead of the reliability of test scores, and so the reliability paradox is a moot issue in item response theory. Third, when the model fits the data, then interpretations of item parameters are not dependent on the characteristics of the examinees (i.e. they are invariant to the latent trait distribution in the population), and the interpretations of person parameters are not dependent on the particular characteristics of the administered items (i.e. they are invariant to the distribution of items). These last two properties make it possible to obtain similar item parameter estimates even when the items are administered to different sets of individuals. Additionally, the similar person estimates can be obtained when different sets of items are administered, respectively. Another advantage is that IRT methods place scores on a common measurement scale and permit the legitimate comparison of latent change over time. Therefore, it is not necessary for individuals to have similar initial status points in order to make comparisons.

4.1 Univariate approaches to repeated measures using IRT

Whenever different forms of the same test are administered (e.g., for test security, to reduce practice effects, for repeated measurement analysis), it becomes necessary to obtain a common metric for the IRT parameters underlying these test forms. By linking the metric of IRT

parameters across test forms, the item parameter scales are corrected for any differences that may arise due to alternative choices for an arbitrary origin and scale unit (Jodoin, Keller, & Swaminathan, 2003). Three common methods for obtaining a common metric are the separate calibration, the concurrent calibration, and the fixed parameter calibration methods. In the separate calibration technique, parameter estimates are first estimated separately for each test, after which, anchor items (i.e. items common in both tests) serve as the basis for the scale transformation to a common metric. However, common items may exhibit drift – a situation where item parameter estimates change in difficulty or discriminating power over time due to outside factors such as increased exposure of the topic (Donoghue & Isham, 1998). In the concurrent calibration process, item parameters are estimated simultaneously by using all responses from the different test forms. Items that are not taken by an examinee are coded as missing. Test forms are linked through the incorporation of anchor items, and the item characteristics of anchor items are assumed to be constant across forms. Ability distributions are allowed to differ in the populations of respondents who receive different test forms (Kim & Cohen, 1998). The fixed parameter calibration method is a variation of concurrent calibration technique; however, the parameters of the common items are treated as known and not estimated. Instead, the item parameters of the remaining items are forced onto the same scale as the common items. Concurrent calibration is the most attractive of the three because it utilizes more information with the potential to provide more accurate estimates (Jodoin, Keller, & Swaminathan, 2003). Regardless, the three methods discussed (as with any unidimensional approach in assessing change) ignore the correlations between latent trait(s) over time, resulting in less precise measurements, especially when tests are short (Wang, Chen, & Cheng, 2004).

4.2 Multivariate Item Response Theory Models for Repeated Measurements

Multidimensional item response theory (MIRT) models have a major advantage over unidimensional item response models for repeated measurements. The estimates of the parameters are more realistic because MIRT models account for the correlations between multiple measures of the latent trait(s) for the same individuals across time. Some suggest (Wang, Chen & Cheng, 2004; Roberts & Ma, 2006) that model parameter estimation is improved when simultaneously calibrating all test items and using these correlations. However, the univariate IRT approaches discussed above ignore such correlations.

4.3 Multidimensional Rasch Model for Learning and Change

Embretson's Multidimensional Rasch Model for Learning and Change (MRMLC; 1991) is a multidimensional IRT model based on a Wiener simplex pattern, where the simplex structure links item responses to an initial ability and one or more modifiabilites that represent the latent changes of individuals between two successive occasions (Embretson, 1991). Belonging to the family of Rasch models, the item discrimination parameter is constrained to one for all items. The MRMLC incorporates the correlations among latent traits between occasions. Additionally, the model directly parameterizes individual change at the latent level.

The MRMLC is given as:

$$P(X_{ni(t)} = 1 \left| \theta^*_{n1}, ..., \theta^*_{nT}, b_i \right.) = \frac{\exp\left(\sum_{q=1}^{t} \theta^*_{nq} - b_i \right)}{1 + \exp\left(\sum_{q=1}^{t} \theta^*_{nq} - b_i \right)}$$

where:

b_i is the difficulty parameter for the ith item;

θ^*_{n1} is the initial latent trait (at occasion t=1) for the nth individual;

θ^*_{n2} is the latent modifiability for the nth individual at occasion 2;

θ^*_{nt} is the latent modifiability for the nth individual at occasion t where $2<t\leq T$;

T is the total number of occasions;

$X_{ni(t)}$ is the nth individual's response to the ith item if and when that item is administered at occasion t with $x_{ni(t)} = 0, 1$.

Under condition t, the latent trait is comprised of the initial ability and t-1 modifiabilities. To estimate the modifiability associated with condition t, item responses across conditions must be combined. The MRMLC model decomposes the ability into an initial latent trait and change in the latent trait across occasions (e.g., θ_2^* represents the change in the latent trait from time 1 to time 2). The use of the model is appropriate when the same items are used across occasions. Different items may also be used across occasions as long as there are some common items across occasions to maintain the metric of the latent trait scale. The MRMLC has been generalized to situations in which a graded polytomous response is obtained. For example, Wang, Wilson, and Adams (1998) developed a partial credit model for repeated measures applications using the same basic idea of parameterizing change as an initial latent trait level and a series of latent change scores. A similar generalized partial credit model for repeated measures (GPCM-RM) was developed by Roberts and Ma (2006).

CHAPTER 5

A NEW IRT MODEL FOR MEASURING CHANGE

A new item response theory model will be presented that allows for the estimation of the latent change scores instead of difference scores, addresses some of the limitations of using difference scores, and provides a direct comparison of the mean latent changes exhibited by different groups (e.g. females versus males). The new item response theory (IRT) model is a multidimensional IRT model that directly parameterizes changes in the latent variable. It generalizes Embretson's (1991) multidimensional Rasch model for learning and change (MRMLC) by allowing item discrimination parameters to vary across items and by estimating a pseudo-guessing parameter for multiple choice items. The model is a type of three-parameter logistic model and can be used with binary data.

The item characteristic function is written as:

$$P(X_{ni(t)}=1\left|\theta^{*}_{n1},...,\theta^{*}_{nT},a_i,b_i,c_i\right.)=c_i+(1-c_i)\frac{\exp\left(a_i\left[\left(\sum_{q=1}^{t}\theta^{*}_{nq}\right)-b_i\right]\right)}{1+\exp\left(a_i\left[\left(\sum_{q=1}^{t}\theta^{*}_{nq}\right)-b_i\right]\right)}$$

where:

a_i is the discrimination parameter for the ith item;

b_i is the difficulty parameter for the ith item;

c_i is the pseudo-guessing parameter for the ith item;

$\theta^{*}_{n1}=\theta_{n1}$ is the initial (i.e., time 1) latent trait level for the nth individual;

$\theta^{*}_{n2}=\theta_{n2}-\theta_{n1}$ is the change in the latent trait from time 1 to time 2 for the nth individual;

$\theta^{*}_{nt}=\theta_{nt}-\theta_{n(t-1)}$ is the change in the latent trait from time t-1 to time t for the nth individual;

$X_{ni(t)}$ is the response of the nth individual to the ith item if it is administered at time t with $X_{ni(t)}$ = 0, 1;

5.1 HYPERPARAMETERS

The θ^*_{nt} variables are assumed to follow a multivariate normal distribution in which two hyperparameters are directly estimated: 1) the centroid and 2) the variance-covariance matrix. The centroid of the θ^*_{nt} parameters is denoted as:

$$\underline{\mu} = [\mu_{\theta^*_1}, \mu_{\theta^*_2}, ..., \mu_{\theta^*_T}]$$

and represents the population mean of each latent variable. The first element of the centroid, $\mu_{\theta^*_1}$, is the average initial status on the latent trait in the population from which the respondents were sampled. Subsequent elements of the centroid provide population averages for the corresponding latent change variables (i.e., the $\theta^*_{\theta_2}, ..., \theta^*_{\theta_T}$ variables). When one of these mean values is close to zero, then one concludes that there has been little change between the corresponding assessment points. On the other hand, values that are substantially positive or negative are indicative of growth or decline between the corresponding assessment points, respectively.

The second hyperparameter is the variance-covariance matrix associated with the θ^*_{nt} variables. As was the case with the centroid, the variance-covariance matrix is estimated simultaneously and directly with the item and person parameters in the model.

The variance-covariance matrix for θ^*_{nt} is denoted as:

$$\sum\nolimits_{\theta} = \begin{pmatrix} \sigma^2_{\theta^*_1} & \sigma_{\theta^*_{12}} & \cdots & \sigma_{\theta^*_{1T}} \\ \sigma_{\theta^*_{21}} & \sigma^2_{\theta^*_2} & \cdots & \sigma_{\theta^*_{2T}} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{\theta^*_{T1}} & \sigma_{\theta^*_{T2}} & \cdots & \sigma^2_{\theta^*_T} \end{pmatrix}$$

The main diagonal of the variance-covariance matrix contains the variance parameters, $\sigma^2_{\theta^*_t}$, that represent variability of the corresponding latent variable. Thus, $\sigma^2_{\theta^*_1}$ represents the variability of the initial latent scores, whereas $\sigma^2_{\theta^*_2}$, …, $\sigma^2_{\theta^*_T}$ represent variances of subsequent latent change scores. The off-diagonal elements contain the covariance parameters for each pair of latent variables. Of particular interest are the off-diagonal elements of the first row (or column) of this matrix in that they represent the linear relationship between initial latent status and latent change between subsequent pairs of assessment points. For example, the literature would predict that $\sigma_{\theta^*_{12}}$ should be negative in that individuals who are above the mean initial status level are expected to exhibit negative latent change at the second assessment point whereas those who are below the mean initial status level are expected to have positive latent change.

If we are interested in examining whether group differences exist in the rate of change, the new model can also be used to estimate the mean latent change trajectory for each group separately. Specifically, the centroid and the variance-covariance matrix associated with the θ^*_{nt} parameters can be estimated separately for each group of interest in a given analysis. By doing so, the new model can be considered a multiple groups model where individuals are from more than one independent population, and group differences in growth trajectories can be statistically tested (e.g. “Do girls learn more algebra than boys during the year on average?”.

This newly developed model is considered multidimensional in form because item responses at later time points depend on more than one latent trait parameter (i.e., an initial latent trait and subsequent latent change variables). However, within any occasion, it is unidimensional in nature because only a single construct is being measured and item response probabilities are given by a unidimensional model for that construct.

5.2 Parameter Estimation and Program Testing

A simulation-based test was conducted to assess the viability of the model parameter estimation. Item response data were generated to mimic the assessment design of the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) for the mathematics subject area only. (This design is thoroughly explained in the following sections of this paper) The ECLS-K is a longitudinal assessment where children are given a two-stage adaptive test at each assessment time (Rock & Pollack, 2002). The assessment design and true item parameters used to simulate item responses were based on previous unidimensional analyses of the ECLS-K mathematics data. The parameters of the new model were estimated using a joint Bayesian solution implemented in the WinBUGS computer program (Speigelhalter et al., 2007). WinBUGS provides a means to conduct Markov Chain Monte Carlo (MCMC) estimation of parameters from a wide variety of models. Because the solutions implemented in WinBUGS are fully Bayesian, prior distributions must be specified for every estimated parameter. The following prior distributions were used for all unconstrained items: 1) log normal (0, .25) distribution of for the discrimination parameters; 2) normal (0, 4) distribution for the difficulty parameters, and 3) beta (6, 16) distribution for the pseudo-guessing parameter. A multivariate normal distribution was used as a prior distribution for θ^{*}_{nt} parameters. The hyperparameters of this distribution (i.e., $\underline{\mu}$ and Σ) were also estimated. A normal (.001, 100) distribution was used to model independently each element of the centroid. A Wishart (I, T) prior distribution was used to estimate elements of Σ, where T is the total number of testing occasions and I is a TxT identity matrix. Spiegelhalter et al. (2007) suggest that this prior distribution is relatively uninformative. The item location and discrimination parameters of one common item were fixed at values reported by NCES (2002) to resolve the indeterminacies in the origin and unit of the latent trait scale. For the initial verification of this estimation strategy, two-stage test data for six assessment points were generated and subsequently analyzed. Ten thousand total MCMC

iterations were performed. Of these iterations, the first 9000 “burn-in” iterations were discarded and the remaining 1000 samples were used to develop expected a posteriori (EAP) estimates of model parameters. The large number of burn-ins was necessary to ensure convergence. However, the necessary number of burn-in iterations was dependent on whether the variance-covariance matrix was allowed to differ across groups. For example, only 9,000 burn-in iterations were necessary when the variance-covariance matrix was constrained to be the same for both groups, while 13,000 burn-ins were necessary when the variance-covariance matrix was allowed to differ between the groups. The test results indicate parameters were estimated accurately using the new model. To ensure convergence with real data, 20,000 burn-ins were used for estimating the parameters in this study, regardless of whether they were constrained or not across gender.

CHAPTER 6

EARLY CHILDHOOD LONGITUDINAL STUDY- KINDERGARTEN COHORT (ECLS-K)

6.1 Description

The Early Childhood Longitudinal Study- Kindergarten Cohort (ECLS-K) is an ongoing longitudinal study conducted by the National Center for Educational Statistics, where students are tested from kindergarten through eighth grade in the areas of science, mathematics, reading, and general knowledge. For this study, the data only include six rounds of testing administered in the fall and spring of kindergarten, fall and spring of first grade, spring of third grade and spring of fifth grade, and only the responses to the mathematical portion were analyzed. For rounds 1-4 (spring and fall of kindergarten and first grade), only one routing form was administered. This form contained 17 items administered to every student. For rounds 5 and 6 (spring of third and fifth grade), two separate routing forms with more difficult items were administered. The forms contained 17 and 18 items, respectively. These three alternate forms contained common items to allow for the establishment of a common metric.

Each round of testing had a two-stage adaptive design. In the first stage, every student was administered the same test form (or routing test). After determining the student's routing test score, he/she was administered the second stage form immediately. The student was given one of three second-stage test forms that were low, moderate, or high in difficulty depending on his /her routing tests score, and some items overlapped between the second-stage forms. The overlap mitigated the emergence of floor or ceiling effects even if a student was placed into the wrong second-stage level. In both stages, the test forms contained common linking items between adjacent testing rounds (Rock & Pollack, 2002). This two-stage adaptive approach allowed for better assessment of both the extent of and variability in growth.

6.2 Mathematical Criterion-Referenced Item Clusters

The math test specifications used in ECLS-K are based on the Mathematical Framework for the 1996 National Assessment of Educational Progress (NAEP; Rock & Pollack, 2002). The specifications are 1) number sense, properties, and operations, 2) measurement, 3) geometry and spatial sense, 4) data analysis, statistics, and probability, and 5) patterns, algebra, and functions. Five specification clusters, assumed to follow a Guttman scale, were developed to identify learning milestones in mathematics by curriculum specialists (Rock & Pollack, 2002). Each cluster contains four items, and a student is considered proficient at any one level if he or she passes any three out of four items. By the eight grade, all clusters were administered to each student.

6.3 Distribution and Number of Items for Each Round of Testing

Table 1 summarizes the number of items for each round of testing for both stages. Some routing and second stage items are common between rounds (See Table 2). For example, four items from the 3rd grade routing test can also be found on the kindergarten/1st grade routing test, and seven items from the 3rd grade routing test are also on the 5th grade low level 2nd stage test. Unique items are those only administered on a specific test form. These items will not be found anywhere else.

Table 1

Total Number of Items in Each Stage for Each Round.

Administered	1st Stage	2nd Stage		
	Routing	Low Level	Moderate Level	High Level
Kindergarten & 1st grade [Rounds 1-4]	17	18	23	31
3rd grade [Round 5]	17	25	21	24
5th grade [Round 6]	18	18	19	24

Table 2

Distribution of Common Items in Other Testing Rounds.

Rounds		Kindergarten & 1st grade				3rd grade				5th grade			
		Route	Low	Mod	High	Route	Low	Mod	High	Route	Low	Mod	High
Kindergarten & 1st grade (8)	**Routing**					4	3	-	-	-	4	-	-
	Low					-	-	-	-	-	-	-	-
	Moderate					-	1	-	-	-	-	-	-
Rounds 1-4	**High**					1	5	-	-	-	-	-	-
3rd grade (11)	**Routing**	4	-	-	1					2	7	2	-
	Low	3	-	1	5					2	-	-	-
Round 5	**Moderate**	-	-	-	-					5	6	2	-
	High	-	-	-	-					6	3	4	5
5th grade (11)	**Routing**	-	-	-	-	2	2	5	6				
	Low	4	-	-	-	7	-	6	3				
Round 6	**Moderate**	-	-	-	-	2	-	2	4				
	High	-	-	-	-	-	-	-	5				

Note: Common items are those found in more than one test form. The number of common routing items found in other test rounds is given in parenthesis.

CHAPTER 7

METHOD

7.1 Participants

In the first part of this study, analyses were performed to examine differential growth between 1000 male and 1000 female randomly selected students. Only 2,000 of the 22,000 students were selected for two reasons. First, only those students with complete data across all 6 rounds were included in this research project. This sample constraint seemed prudent given the desire to avoid mischaracterizations of the model at this early stage of model development. When considering only those students with complete data, the effective respondent pool decreased to approximately 2,500 individuals. The sample size used in this part of the study was further reduced to 2000 due to the computational limitations of WinBUGS, such as processing speed and limited memory capacity. In the second part of this study, analyses were performed to examine differential growth in mathematical ability between 1,443 Caucasians and 282 African American students. The sample size was smaller than the gender analysis, and group sizes were not equal because in addition to taking all six rounds of testing, ethnicities had to be reported.

7.2 Model variations

Six variations of the model were investigated by modifying the constraints placed on Σ_θ, $\underline{\mu}$, or both. Table 3 summarizes the different models that were studied. Each model was run using responses from 2,000 students in the gender analyses and 1,725 students for the ethnicity analyses. Models 1, 3, and 5 are similar in that Σ_θ is constrained to be equal across groups, but the models differ with regard to the constraints placed on the centroid. In model 1 the centroid is constrained to be equal across groups whereas in model 5 the centroids are estimated separately for each group. In model 3 the centroids are free to differ at baseline but constrained to be equal across groups for subsequent time points. Models 2, 4, and 6 are

similar in that Σ_θ is estimated separately for each group but differ on the constraints placed on the centroid. The same constraints placed on the centroid for models 1, 3, and 5 are placed on models 2, 4, and 6, respectively.

Table 3

Variations of the New IRTMmodel

Variations of the Model	Σ_θ constrained to be equal across groups	Σ_θ estimated for both groups separately
Centroid constrained to be equal across groups	Model 1	Model 2
Centroid allowed to differ at baseline but constrained to be equal afterwards	Model 3	Model 4
Centroid estimated for both groups separately	Model 5	Model 6

7.3 Convergence and model selection using deviance information criterion (DIC)

Twenty-one thousand iterations were used to ensure convergence, where 20,000 iterations served as burn-in and the last 1,000 iterations were used to develop EAP estimates of model parameters. Additionally, trace plots of all item parameters and hyperparameters were also examined to ensure all variations of the model converged (see Figures 15-16). Evidence for convergence is provided when segments of the trace plot for a parameter traverse the same parts of the sample space and no clear pattern is found (e.g. always decreasing) (Sinharay, 2003). The trace plots of centroid parameters for both gender and ethnicity results can be seen in figures 15 and 16, respectively. The deviance information criterion (DIC; Speigelhalter, Best, Carlin & Van der Linden, 2002) was then used to select the model that best fit the ECLS-K data. The DIC takes into account the model fit and the complexity of the model which is measured by the effective number of parameters that are estimated. As the number of parameters increases the DIC value also increases. Models with smaller DIC values are favored. The DIC values were very similar for all six models; however, the smallest DIC value (DIC= 414,000 and

DIC=356,942 for gender and ethnicities, respectively) resulted when Σ_θ and $\underline{\mu}$ were not constrained (i.e., when both sets of parameters were estimated separately in each group). Therefore, model 6 was selected for both the gender and ethnicity analyses. Table 4 and Table 5 summarize the DIC for each variation of the model for the gender and ethnicity results, respectively.

Table 4

Gender: DIC Values for Each Variation of the Model

Variations of the Model	Σ_θ constrained to be equal across groups	Σ_θ estimated for both groups separately
Centroid constrained to be equal across groups	414,026	414,014
Centroid allowed to differ at baseline but constrained to be equal afterwards	414,062	414,055
Centroid estimated for both groups separately	414,040	414,000

Table 5

Ethnicity: DIC Values for Each Variation of the Model

Variations of the Model	Σ_θ constrained to be equal across groups	Σ_θ estimated for both groups separately
Centroid constrained to be equal across groups	357,037	356,991
Centroid allowed to differ at baseline but constrained to be equal afterwards	356,948	356,992
Centroid estimated for both groups separately	356,978	356,942

7.4 Item fit

Item responses and item parameter estimates were used to plot mean observed versus expected proportions within homogeneous θ groups. Correlations between the average observed and average expected values for each item were also calculated. Using θ estimates

obtained from model 6, respondents were sorted and homogeneous groups of approximately equal size ($N \approx 50$) were formed. Within each group, the average θ, the average observed proportion of examinees who answered a given item correctly, and the expected proportion of correct responses were calculated. These mean observed and expected responses were then plotted by average θ and also correlated across θ groups. The correlation can be used as a measure of item fit that is sensitive to the scatter of observed data relative to its expected value (It is less sensitive to misfitting trend). For binary data, experience suggests that correlations less than .9 are indicative of items with too much scatter. All items were examined graphically to assess the fit of the model separately for gender and ethnicity. Most items had high correlations ($r > .90$) and were well-fitting. Therefore, only items with extremely high correlations and items suspect of misfit with respect to either trend or scatter are depicted here to conserve space. In each of the figures that follow, the top panel illustrates a well-fitting item, whereas the bottom panel depicts the most misfitting item within an item set of interest.

7.4.1 Gender: common items

Four items common across rounds 1-6 and nine items common across rounds 1-4 and round 5 have high correlations ranging from .95 to .99, and no items exhibited problems with misfitting trend or scatter. Items 15 ($r = .995$) and 13 ($r = .990$) represent the best and worst fitting items, respectively, among the four common items across rounds 1-6 (see Figure 17). Items 8 ($r = .993$) and 48 ($r = .950$) represent the best and worst fitting items, respectively, among the nine items common across rounds 1-4 and round 5 (see Figure 18). Though considered the worst fitting items, both items 13 and 48 have reasonably good fit. Twenty-seven common items were administered in round 5 and round 6, and correlations ranged from .85 to .98. Item 73 had the highest correlation ($r = .993$) and is depicted in the top panel of Figure 19. The fit of item 122 ($r = .85$) was suspicious due to its low correlation (see Figure 19-bottom panel); however, a closer inspection revealed only a slight problem with scatter.

7.4.2 Gender: unique items

Fifty-one unique items were administered during rounds 1-4, and only three items had correlations less than .90. Item 4 and item 18 had the highest (r = .990) and lowest correlation (r = .67), respectively (see Figure 20). Though the correlation is low, scatter does not appear to be a problem for item 18. Instead, attenuation of the correlation appears to result from a lack of response variability for this item (i.e., a ceiling effect). Thirty-four and twenty-nine unique items were administered during round 5 and round 6, respectively, and most items had high correlations. The two of the best fitting items were item 77 (r = .989) and item 22 (r = .988) for round 5 and round 6, respectively (see top panels of Figures 21 and 22). Upon inspection of the eight suspicious items in round 5 and two suspicious items in round 6, all but two items appeared to fit visually and did not exhibit obvious misfitting trend or extreme scatter. Item 117 (r = .68) and item 205 (r = .85) exhibited more pronounced scatter (see bottom panels of Figure 21 and Figure 22).

7.4.3 Ethnicity: common items

The four items administered in the six rounds of testing and seven of the nine items administered during rounds 1-4 and round 6 had extremely high correlations (r > .98); while the remaining two items common across rounds 1-4 and round 6 have high correlations (r=.96). Among the four common items across rounds 1-6, the best and worst fitting items were item 15 (r = .995) and item 13 (r = .989), respectively (see Figure 23). Among the nine items common across rounds 1-4 and round 6, item 52 (r = .990) and item 48 (r = .958) had the highest and lowest correlations, respectively (see Figure 24). Of the twenty-seven items common across round 5 and round 6, seven items had correlations lower than .90 but most items did not exhibit obvious scatter. Item 73 had the highest correlation of .983, and item 116 had the lowest correlation and a closer inspection revealed a problem with scatter (see Figure 25).

7.4.4 Ethnicity: unique items

One hundred fourteen unique items were administered in rounds 1-6, and most items had high correlations. Items 7 (r = .99), 74 (r = .983) and 178 (r = .986) had the highest correlations among the unique items in rounds 1-4, round 5, and round 6, respectively (see top panels of Figures 26-28). Upon inspection of the seven suspicious items from rounds 1-4, eight suspicious items in round 5, and five suspicious items in round 6, all items appeared to fit visually and did not exhibit obvious misfitting trend or scatter. The following items were suspect of scatter but appear to fit visually: Item 18 (r = .66), item 117 (r = .70), and item 205 (r = .872) from rounds 1-4, round 5, and round 6, respectively, can be seen in the bottom panels of Figures 26-28.

These initial analyses suggest the model fits both the gender and ethnicity data.

CHAPTER 8

RESULTS/ DISCUSSION

8.1 Gender: Item parameter estimates

A total of 153 unique items were administered across all six rounds; however, the number of item responses to the repeated assessments varied given the nature of the two-stage adaptive testing design. Recall that the same test forms were administered for rounds 1 through 4. The parameters (b=.35; a=1.8) of item 13 were constrained to resolve the indeterminacies in the origin and unit of the latent trait scale, and this item was common across all six rounds. Appendix A lists the item discrimination (a_i), item difficulty (b_i), and pseudo-guessing parameter (c_i) estimates for the common items. The estimated parameters for the unique items for rounds 1-4, round 5, and round 6 are listed in Appendices B-D, respectively. Unique items from rounds 1-4 had extremely low difficulty parameter estimates ranging from -5.07 to -8.62 (see Appendix B). Upon examination, four of these items were found to be low 2nd stage test items from rounds 1-4 and the remaining item was the easiest routing item for those rounds. Ignoring these five items, item difficulty estimates ranged from -4.58 to+ 2.588 for rounds 1-4. Item difficulty estimates for round 5 and round 6 unique items ranged from -2.024 to +3.400 and +1.413 to +4.383, respectively,(see Appendices C-D). Most items had moderate to high levels of discrimination, and 15 items had low levels of discrimination where a_i was less than one. Ten of these items were administered during rounds 1-4, where items were generally easier, making it harder to discriminate among the individuals. The remaining five items were administered during round 5. All items in round 6 had discrimination parameters greater than 1.3, Of the 34 pseudo-guessing parameters estimated for the multiple choice items, only one item from round 6 had a pseudo-guessing parameter estimate less than 0.1. Thus, the results indicated that a three-parameter model was necessary for multiple choice items in this dataset.

The average item parameter estimates and standard deviations for each round and stage of testing are reported in Table 6. The average difficulty increased across the second stages within each round of testing as well as among rounds. The average discrimination levels increased across each round of testing; however, within each round, no general trend was found. The average pseudo-guessing parameter estimates decreased across and within each round testing.

Table 6

Gender: Average Item Parameter Estimates and Standard Deviations for Each Round and Stage of Testing

Round	Stage Level	a_i		b_i		c_i	
		Mean	STD	Mean (std)	STD	Mean (std)	STD
1-4	Routing	1.616	0.435	-1.876	2.147	0.212	0.069
	Low	1.137	0.377	-3.599	1.944	0.284	0.086
	Moderate	1.518	0.868	-1.830	1.264	0.212	0.062
	High	1.665	0.419	-0.267	1.603	0.184	0.060
5	Routing	2.007	0.362	1.423	0.696	0.127	0.011
	Low	1.712	0.733	0.208	1.012	0.218	0.033
	Moderate	1.906	0.735	1.362	0.717	0.182	0.050
	High	1.876	0.552	2.539	0.681	0.139	0.035
6	Routing	2.191	0.583	2.249	0.762	0.165	0.059
	Low	1.821	0.621	1.321	0.698	0.147	0.051
	Moderate	1.788	0.579	2.506	0.695	0.143	0.059
	High	2.039	0.481	3.346	0.558	0.127	0.025

The item characteristic curves (ICC) of four common items and 9 unique second stage low, moderate, and high difficulty level items for each round of testing are portrayed in figures 1-4, respectively. Items representative of the typical item parameter estimates for each round and stage of testing were chosen to display here. The ICC portrays the probability of correctly responding to that item as a function of the composite theta (θ_{nt}). In Figures 2-4 note that as the

item difficulty increases, discrimination also increases (i.e., the curves shift to the right and the slopes get steeper).

8.2 Gender: Test Information Function

Figures 5-8 illustrate the test information function (TIF) for rounds 1-4, round 5 and round 6 items, respectively, for each stage and form of testing. The curves portray test information as a function of the composite theta (θ_{nt}). In Figure 5, the routing test for rounds 1-4 provided the most information (i.e., precision for estimating ability) at $\theta_{nt} = 0.15$, while the low level test form of the second stage provided the most amount of information at $\theta_{nt} = -4.48$. Note that the TIF curve of the low test form is very flat relative to the other curves in the figure, and among all test forms at a given stage (except for Round 6), its maximum TIF is the smallest (see Table 7). This can be expected due to the extremely low level of difficulty (i.e., the easiest of all test forms) and the low discriminating power of the items which makes it harder to differentiate among the moderate to high ability levels. Generally, across and within each round of testing (see Figures 5-7), the maximum TIF and the location of maximum information (θ_{nt}) increase with test difficulty due to the increase in the number of more discriminating items; however, for round 5 there were more discriminating items on the moderately difficult test form than the high difficulty test form, resulting in a higher maximum TIF than on the high difficulty test form. Also, this trend is not true for round 6, probably due to the lower number of discriminating items on the moderately difficult test form as compared to the low difficultly and high difficulty test forms. Table 7 summarizes the values and locations of the maximum test information function for all rounds of testing.

Table 7

Gender: Location of Maximum Test Information Function.

Administered	1st Stage			2nd Stage								
	Routing			Low Difficulty			Moderate Difficulty			High Difficulty		
	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF
Kindergarten & 1st grade [Rounds 1-4]	17	0.15	4.94	18	-4.48	2.51	23	-1.61	13.05	31	0.53	14.01
3rd grade [Round 5]	17	1.31	12.59	22	1.09	13.60	24	1.36	20.02	23	2.3	14.51
5th grade [Round 6]	18	2.12	13.69	18	1.5	11.74	19	2.28	9.84	20	3.39	15.52

8.3 Gender: Centroid

Direct estimates of the centroid (population mean for each θ_{nt}^{*}) for both groups were obtained using the model. Recall that θ_{n1}^{*} is the initial mathematical ability of each student in the fall of kindergarten and subsequent θ_{nt}^{*} elements represent the change (or growth) in mathematical ability from one assessment period to the next. The average baseline ability for the population is represented by $\hat{\mu}_1$ and found to be -2.724 and -2.573 for females and males, respectively. These baseline ability estimates were standardized to center the scale at 0 by subtracting the combined average of the two groups and then dividing by the pooled variance calculated from the latent trait estimates within each group. Estimated mean levels of change are reported in Table 8, and these estimates were standardized to the same baseline metric. The standardized estimates for $\hat{\mu}_1$ were -.06 and .07 for females and males, respectively, and a two-tailed t-test was conducted to test for group differences and found to be statistically significant (t= -2.60, p<.05). However, the difference between the two groups was only a .13 standard deviation of the initial mathematical ability level and may be too small to have pragmatic implications for practitioners.

The standardized estimated mean levels of change in mathematical ability from fall to spring of kindergarten ($\hat{\mu}_2$ =1.04) and from spring of kindergarten to fall of 1st grade ($\hat{\mu}_3$ =.51) were equal for both groups and nearly equal for Round 4 from fall to spring of 1st grade. For the first two years of education, the largest amount of growth occurred between fall and spring of each year and growth rates were nearly identical for male and females. Differences in growth between the two groups were not found until the spring of 3rd grade and the spring of 5th grade, where $\hat{\mu}_5$ was equal to 1.14 and 1.20 and $\hat{\mu}_6$ was equal to 0.72 and 0.75 for females and males, respectively. Though there was only a .06 standard deviation difference between the groups, $\hat{\mu}_5$ was found to be statistically different (t= -2.48, p<.05) between males and females but such

a small difference may have little substantive meaning. No statistical differences were found for round 6 (t= -1.10, p>.05).

Table 8

Gender: Estimated Mean Levels of Change in Mathematical Ability.

	Unstandardized		Standardized	
	Females	Males	Females	Males
$\hat{\mu}_1$ (baseline)	-2.72	-2.57	-0.06	0.07
$\hat{\mu}_2$	1.19	1.19	1.04	1.04
$\hat{\mu}_3$	0.59	0.58	0.51	0.51
$\hat{\mu}_4$	1.16	1.15	1.01	1.00
$\hat{\mu}_5$	1.31	1.38	1.14	1.20
$\hat{\mu}_6$	0.82	0.86	0.72	0.75

8.4 Gender: Variance-covariance matrix

Using the model, direct estimates of the variance-covariance matrix for θ^*_{nt} measures were obtained to examine variability in baseline ability and growth for both groups. Recall that the first element in the main diagonal of ∑ represents the variability in mathematical ability at baseline, and the subsequent main diagonal elements represent the variability of change in ability. Only elements from the main diagonal of ∑ (i.e., $\sigma^2_{\theta^*}$) are reported in Table 9 for each group:

Table 9

Variability in Baseline Ability and Growth for Females and Males.

	Females	Males
$\sigma^2_{\theta_1^*}$ (baseline)	1.382	1.441
$\sigma^2_{\theta_2^*}$	0.243	0.290
$\sigma^2_{\theta_3^*}$	0.179	0.207
$\sigma^2_{\theta_4^*}$	0.251	0.259
$\sigma^2_{\theta_5^*}$	0.183	0.225
$\sigma^2_{\theta_6^*}$	0.099	0.097

Females and males have similar variability in baseline ability and growth, and the largest difference between the two groups was only .059. The variability of growth fluctuated, decreasing and increasing, between assessment points until the spring of 5th grade where the variability was only .099 and .097 for females and males, respectively; indicating the variability in growth rate was similar for these two groups.

The estimated variance-covariance matrix for θ^*_{nt} were converted into a correlation matrix for each group:

$$\hat{P}_{\theta^*_{Females}} = \begin{pmatrix} 1.00 & -.45 & -.23 & -.46 & -.10 & .21 \\ -.45 & 1.00 & -.21 & .03 & .02 & -.02 \\ -.23 & -.21 & 1.00 & -.19 & -.02 & -.08 \\ -.46 & .03 & -.19 & 1.00 & -.28 & -.08 \\ -.10 & .02 & -.02 & -.28 & 1.00 & -.13 \\ .21 & -.02 & -.08 & -.08 & -.13 & 1.00 \end{pmatrix}$$

$$\hat{P}_{\theta^*_{Males}} = \begin{pmatrix} 1.00 & -.41 & -.14 & -.45 & -.22 & .06 \\ -.41 & 1.00 & -.26 & .07 & .01 & -.01 \\ -.14 & -.26 & 1.00 & -.22 & .00 & .01 \\ -.45 & .07 & -.22 & 1.00 & -.25 & .04 \\ -.22 & .01 & .00 & -.25 & 1.00 & -.15 \\ .06 & -.01 & .01 & .04 & -.15 & 1.00 \end{pmatrix}$$

The off-diagonal elements on the first row of both matrices indicated that there was a low to moderate negative correlation between initial ability and subsequent changes except in round 6 where there was a slight positive correlation. This suggested that, on average, individuals who were below the mean at baseline experienced positive growth (relative to the mean) until the spring of fifth grade where they experienced negative growth (i.e., the amount of change was below the mean of the group).

8.5 Ethnicity: Item parameters estimates

The estimated item parameters were similar to those of the gender results (see Table 10). Within each round of testing, the average item difficulty levels increased with each secondary stage of testing (i.e., low, moderate, and high), and in general, the average discrimination levels also increased; however, in round 6, the average discrimination level was lower for the moderately difficult test form than the low difficulty test form.

Table 10

Ethnicity: Average Item Parameter Estimates and Standard Deviations for Each Round and Stage of Testing

Round	Stage Level	a_i		b_i		c_i	
		Mean	STD	Mean (std)	STD	Mean (std)	STD
1-4	Routing	1.546	0.417	-2.013	2.268	0.214	0.059
	Low	1.070	0.326	-3.809	2.138	0.296	0.089
	Moderate	1.513	0.847	-1.878	1.319	0.230	0.056
	High	1.651	0.419	-0.291	1.646	0.214	0.084
5	Routing	1.945	0.342	1.399	0.699	0.133	0.021
	Low	1.690	0.729	0.119	1.047	0.225	0.034
	Moderate	1.835	0.734	1.302	0.843	0.176	0.048
	High	1.847	0.563	2.550	0.708	0.135	0.039
6	Routing	2.149	0.575	2.251	0.778	0.171	0.071
	Low	1.785	0.599	1.303	0.680	0.143	0.073
	Moderate	1.747	0.580	2.491	0.695	0.129	0.058
	High	2.019	0.503	3.361	0.592	0.128	0.032

The item characteristic curves (ICC) of nine second stage low, moderate, and high difficulty level items for each round of testing are portrayed in figures 8-10, respectively. These nine items represent the prototypical item parameter estimates for each round and stage of testing. Note that as the item difficulty increases across the rounds, discrimination also increases (i.e., the curves shift to the right and the slope becomes steeper). This is most noticeable when contrasting Rounds 1-4 with later rounds.

8.6 Ethnicity: Test Information Function

Figures 11-13 illustrate the test information function (TIF) for Rounds 1-4, Round 5 and Round 6 items, respectively, for each stage and form of testing. In Figure 11 note that the TIF curve of the low test form in rounds 1-4 is the flattest relative to the other curves in the figure. Among all test forms at a given stage, its maximum TIF is the smallest but provides much more information in later rounds than in rounds 1-4 (see Table 11). The low discriminating power of

the items and the extremely low level of difficulty of the low difficulty test form make it harder to differentiate among the moderate to high ability levels. Generally, across and within each round of testing (see Figures 11-13), the maximum TIF and the location of maximum information (θ_{nt}) increase with test difficulty due to the increase in the number of more discriminating items; however, as seen with the gender results, this trend is not true for rounds 5 and 6. There were more discriminating items on the moderately difficult test form than the high difficulty test form in round 5, resulting in a higher maximum TIF than on the high difficulty test form. In round 6, the maximum TIF value may be lower due to the lower number of discriminating items on the moderately difficult test form as compared to the low difficultly and high difficulty test forms. Table 11 summarizes the values and locations of the maximum test information function for all rounds of testing.

Table 11

Ethnicity: Location of Maximum Test Information Function.

Administered	1st Stage			2nd Stage								
	Routing			Low Difficulty			Moderate Difficulty			High Difficulty		
	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF	# of items	θ_{nt}	TIF
Kindergarten & 1st grade [Rounds 1-4]	17	0.11	4.72	18	-4.38	1.93	23	-1.64	12.49	31	0.46	14.02
3rd grade [Round 5]	17	1.26	11.85	22	1.04	13.14	24	1.35	19.15	23	2.30	14.24
5th grade [Round 6]	18	2.07	13.41	18	1.49	11.39	19	2.22	9.50	20	3.46	15.11

8.7 Ethnicity: Centroid

Estimates of the centroid were obtained separately for each group, and then standardized in the same manner as in the gender analyses (See Table 9). The standardized estimates for $\hat{\mu}_1$ were 0.38 and -.38 for Caucasians and African Americans, respectively, and these results indicate that African American students are 0.76 standard deviations behind Caucasians in their the initial mathematical ability level. However, African American students have slightly larger growth rates until spring of 5th grade but a statistical difference in θ* was found only at round 3 with a difference of .08 standardized deviation units between the groups. Substantively this difference may not be important. Growth trajectories (the accumulation of standardized μ_t) for Caucasians and African Americans were plotted (see Figure 14). It is apparent that large differences exist in mathematical ability in the fall of kindergarten but remain relatively constant over time. Growth for both groups accelerated over time except in the fall of first grade and spring of fifth grade, where growth was smallest, yet still quite evident.

Table 12

Ethnicity: Estimated Mean Levels of Change in Mathematical Ability.

	Unstandardized		Standardized	
	African Americans	Caucasians	African Americans	Caucasians
$\hat{\mu}_1$ (baseline)	-3.40	-2.44	-0.38	0.38
$\hat{\mu}_2$	1.29	1.21	1.04	0.98
$\hat{\mu}_3$	0.67	0.57	0.54	0.46
$\hat{\mu}_4$	1.18	1.12	0.95	0.91
$\hat{\mu}_5$	1.37	1.34	1.10	1.08
$\hat{\mu}_6$	0.80	0.89	0.64	0.71

8.8 Ethnicity: Variance-covariance matrix

Variability in baseline ability and growth for both groups was examined using the variance-covariance matrix for θ^*_{nt} measures obtained directly by the model. The $\sigma^2_{\theta^*}$ are reported in Table 13 for each group:

Table 13

Variability in Baseline Ability and Growth for African Americans and Caucasians.

	African Americans	Caucasians
$\sigma^2_{\theta_1^*}$ (baseline)	1.454	1.304
$\sigma^2_{\theta_2^*}$	0.248	0.294
$\sigma^2_{\theta_3^*}$	0.212	0.196
$\sigma^2_{\theta_4^*}$	0.326	0.215
$\sigma^2_{\theta_5^*}$	0.227	0.191
$\sigma^2_{\theta_6^*}$	0.011	0.095

African Americans and Caucasians had similar variability in baseline ability and growth, and the largest difference between the two groups was .15 at baseline. Generally, the variability of growth was similar for both groups and decreased over time except in the spring of 1st grade there was greater variability among the African American students than the Caucasian students. However, by the spring of 5th grade variability was very small for both groups.

The estimated variance-covariance matrix for θ^*_{nt} were converted into a correlation matrix for each group:

$$\hat{P}_{\theta^*_{AA}} = \begin{pmatrix} 1.00 & -.46 & -.28 & -.45 & -.15 & -.06 \\ -.46 & 1.00 & -.07 & .03 & -.08 & -.18 \\ -.28 & -.07 & 1.00 & -.26 & .10 & .05 \\ -.45 & .03 & -.26 & 1.00 & -.19 & .10 \\ -.15 & -.08 & .10 & -.19 & 1.00 & -.07 \\ -.06 & -.18 & .05 & .10 & -.07 & 1.00 \end{pmatrix}$$

$$\hat{P}_{\theta^*_{Caucasians}} = \begin{pmatrix} 1.00 & -.49 & -.14 & -.46 & -.16 & .16 \\ -.49 & 1.00 & -.32 & .15 & -.03 & -.06 \\ -.14 & -.32 & 1.00 & -.27 & -.01 & .01 \\ -.46 & .15 & -.27 & 1.00 & -.20 & -.05 \\ -.16 & -.03 & -.01 & -.20 & 1.00 & -.13 \\ .16 & -.06 & .01 & -.05 & -.13 & 1.00 \end{pmatrix}$$

The off-diagonal elements on the first row of the matrix indicated that there was a low to moderate negative correlation between initial ability and subsequent changes for African Americans; This suggested that, on average, individuals who were below the mean at baseline experienced positive growth (relative to the mean) throughout their elementary education. The same was true of Caucasians except in round 6 where there was a slight positive correlation. Thus, Caucasians who were above the mean at baseline generally experienced below average growth from spring of 3rd grade to spring of 5th grade.

CHAPTER 9

CONCLUSION

9.1 Gender and Ethnicity

In educational research students are tested repeatedly to asses their level and change in subject area knowledge, and additionally, group differences (e.g. gender or racial differences) are often examined. In the education literature, the findings are inconsistent with regard to when gender differences manifest in mathematical ability, the magnitude of these differences and why they occur. Some authors suggest that performance differences manifest as early as elementary school, while others suggest middle school (Hyde, Fennema, & Lamon, 1986). Aunola, Leskinen, Lerkkanen, and Nurmi (2004) suggest that gender differences do not exist, but rather it is the initial status that determines the development of mathematics proficiency, which may proceed in one of two ways: 1) children who start with good skills have a greater change in proficiency (i.e., learn more) than those who don't; 2) children who originally start with a low level of skills and related knowledge increase the speed of their development and catch up with those who originally have higher levels of these. In Hyde, Fennema, & Lamon's (1986) meta-analysis of 100 studies, the authors concluded that gender differences in mathematical performances are small. Regardless of these inconsistencies, however, the general consensus among researchers are 1) male students outperform female students and 2) Caucasian students outperform African American students in mathematics.

A new item response theory model was used to examine differential growth in mathematical ability between male and female students as well as Caucasians and African American students from kindergarten through fifth grade. Model 6 (where Σ and $\underline{\mu}$ were estimated separately for each group) was selected to examine differences in mean growth for both gender and ethnicity. The findings suggest that differential growth may not exist between male and female students. Growth rates were identical for the first two years of schooling for

male and female students. Additionally, students experienced the largest amount of growth from fall of kindergarten to spring of first grade. Growth may not have been as substantially large between spring of kindergarten to fall of first grade due to drop in knowledge retention during the summer break. Though there are statistical differences (favoring the males) in baseline mathematical ability and growth for one round of testing, these differences may be too small to have substantive meaning. Some dissimilarities were found in the variability of baseline ability and growth between the two groups, but again, these were not substantial. This suggests that the DIC can be very sensitive to detecting rather small group differences that may not meaningful in educational practice.

The ethnicity findings demonstrate that growth rates were similar for Caucasians and African Americans; however, average baseline ability in mathematics differed substantially in these samples. The results indicated that as early as fall of kindergarten, African Americans are 0.76 standard deviations behind their cohorts in their initial mathematical ability. With such a gap existing at the start of kindergarten and growth being nearly equal between the groups over the years, achievement gaps between the groups are not reduced by elementary education. While African Americans are making achievement gains during this period, these gains may be on low-level and basic mathematic skills (Tate, 1997). Additionally, mathematics is a hierarchically arranged subject, with each step drawing upon knowledge and skills from the preceding step, so differences in skill development in lower grades may set barriers to acquisition of more complex skills needed to succeed. It is estimated that half of the achievement gap between races in middle and high school can be accounted by the differences detected at the kindergarten level (NCES, 1995). Given that these differences are apparent at the onset of formal schooling, timing of educational intervention in schools and families may need to occur much earlier. Programs such as Head Start attempt to address these needs for low-income, pre-school aged children.

In studying and better understanding the changes in proficiency at the individual level and identifying when these changes occur, teaching practices can be targeted to foster equity in mathematics, or any other subject. The new model can address these issues while providing more precise estimates of change than those derived by NCES (2002) with a unidimensional model by directly comparing the mean latent changes exhibited in different groups.

9.2 Limitations and future research

There are several limitations to this study. First, only a subset of the entire ECLS-K dataset was used for three reasons: 1) WinBUGS did not have the capacity to run with the entire dataset, 2) WinBUGS is extremely slow, taking as long as 10 days to analyze data for one model using a single MCMC chain, and 3) only participants who responded during all six rounds of assessment were included. Second, to simplify initial model development the study assumed that change in mathematical ability is unidimensional. This is a common assumption used to measure change in mathematical ability. However, the justification for this assumption is an empirical issue that remains to be studied in the future. For example, Roberts and Ma (2006) and te Marvelde et al. (2006) have both suggested models in which general constructs like mathematics can change along several specific dimensions. The technical and test design hurdles associated with the application of these models will be quite high, but may be matched by improved insight about individual changes in mathematics ability. Future research is necessary and will attempt to address these limitations. First, model specific MCMC programs must be developed in a primary computing language (e.g., FORTRAN, C++) in order to increase the speed and efficiency of parameter estimation. Second, a new extension of this model called the “Sprout Model” (Roberts & Ma, 2006) is currently under development. The Sprout Model is designed theoretically designed to assess growth in a truly multidimensional space in which the existence of specific dimensions can vary somewhat over time.

Appendix A

Common Items

Common across ALL Rounds			
Items	**a**	**b**	**c**
13	1.800	0.350	0.000
14	1.899	0.357	0.000
15	1.909	0.779	0.000
16	1.961	0.571	0.000

Common across Rounds 1-4 & Round 5			
Items	**a**	**b**	**c**
11	1.162	-1.115	0.000
38	1.673	-0.802	0.000
12	1.654	-0.678	0.000
8	1.829	-0.425	0.000
60	1.896	0.409	0.000
49	1.289	0.844	0.000
51	2.148	0.905	0.000
48	1.013	0.918	0.000
52	1.994	1.189	0.000

Common across Round 5 & Round 6							
Items	**a**	**b**	**c**	**Items**	**a**	**b**	**c**
116	0.624	0.334	0.000	**129**	1.860	2.052	0.000
113	2.085	0.933	0.000	**80**	2.647	2.161	0.119
107	1.954	0.990	0.000	**78**	1.814	2.322	0.000
70	1.746	1.069	0.000	**127**	2.232	2.399	0.000
106	3.143	1.358	0.000	**125**	1.753	2.601	0.000
108	1.012	1.390	0.000	**141**	2.191	2.679	0.103
102	2.538	1.402	0.000	**121**	0.414	2.798	0.000
73	2.296	1.483	0.000	**81**	2.228	2.803	0.000
76	1.664	1.487	0.000	**134**	1.433	2.811	0.000
118	3.062	1.532	0.000	**138**	2.108	3.008	0.000
115	1.822	1.555	0.000	**135**	0.878	3.176	0.000
120	2.406	1.739	0.116	**139**	1.970	3.182	0.000
122	1.827	1.995	0.000	**140**	1.904	4.534	0.000
126	2.414	2.023	0.000				

Appendix B

Unique Items in Rounds 1-4 Testing

Items	a	b	c	Items	a	b	c
17	0.981	-8.624	0.000	**26**	0.898	-1.801	0.316
64	2.674	-7.740	0.000	**39**	3.833	-1.719	0.000
18	0.559	-6.123	0.446	**42**	4.311	-1.575	0.000
23	0.728	-5.882	0.322	**41**	1.228	-1.537	0.000
20	1.932	-5.079	0.000	**30**	1.018	-1.369	0.000
28	0.613	-4.583	0.000	**10**	1.690	-1.125	0.136
21	1.853	-4.375	0.000	**40**	1.686	-1.111	0.000
33	1.078	-3.925	0.000	**37**	1.208	-1.095	0.114
6	0.850	-3.841	0.229	**36**	0.919	-1.038	0.170
2	0.945	-3.722	0.000	**29**	1.355	-0.907	0.000
3	1.796	-3.692	0.000	**43**	1.686	-0.391	0.000
32	1.064	-3.595	0.000	**59**	2.055	-0.221	0.000
31	1.261	-3.407	0.219	**58**	2.178	-0.166	0.000
19	0.856	-3.131	0.000	**61**	1.845	-0.056	0.000
24	1.332	-2.825	0.275	**47**	1.742	0.072	0.000
1	1.340	-2.674	0.272	**44**	1.417	0.121	0.000
22	1.526	-2.652	0.000	**57**	1.585	0.246	0.000
25	0.949	-2.495	0.198	**45**	2.041	0.523	0.000
5	1.540	-2.309	0.000	**62**	2.160	0.541	0.000
4	1.640	-2.274	0.000	**63**	2.267	0.644	0.000
9	1.200	-2.204	0.000	**50**	2.385	0.965	0.000
7	1.575	-2.142	0.000	**55**	1.849	1.428	0.000
46	1.894	-2.116	0.000	**54**	2.180	1.519	0.000
34	1.228	-2.003	0.219	**56**	1.287	1.977	0.000
27	1.233	-2.002	0.000	**53**	1.911	2.588	0.000
35	1.363	-1.909	0.188				

Appendix C

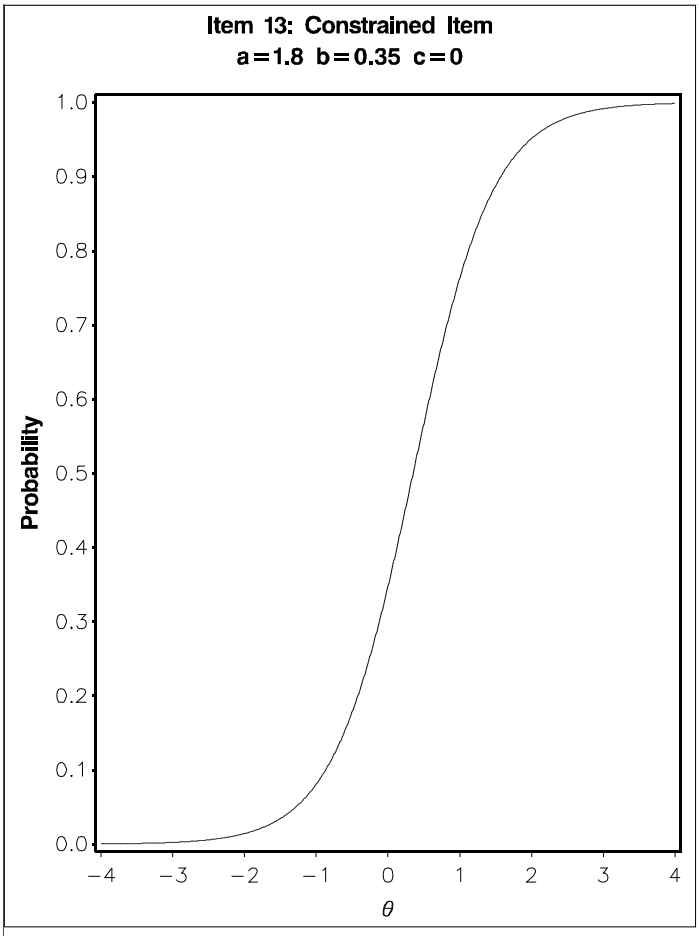
Unique Items in Round 5 Testing

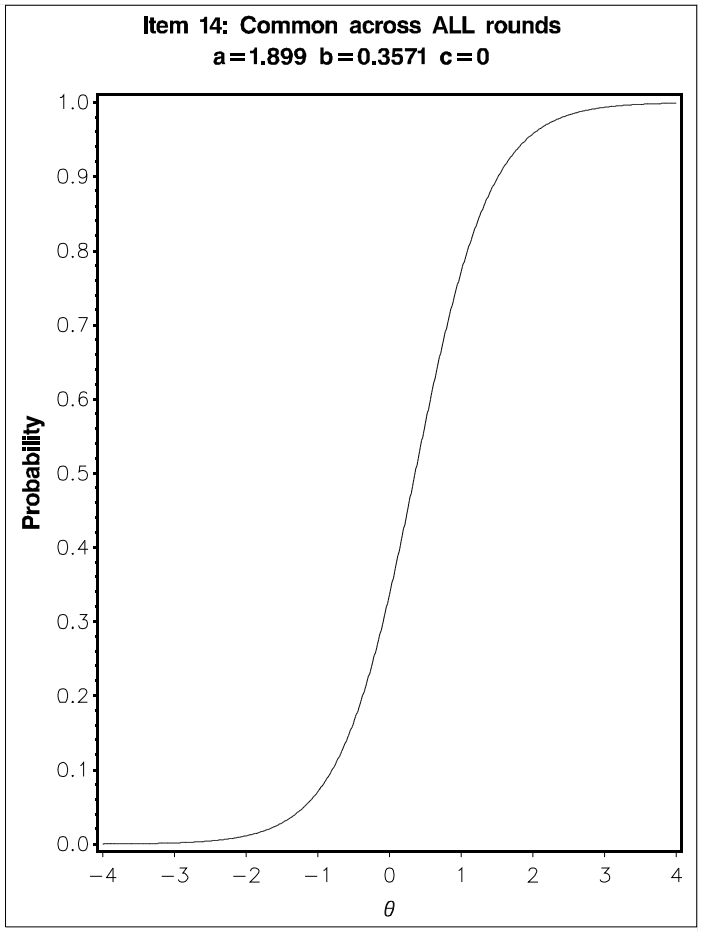
Items	a	b	c
92	0.978	-2.024	0.000
90	1.474	-1.761	0.000
88	0.812	-0.605	0.000
83	1.223	-0.329	0.255
109	1.888	-0.271	0.000
97	0.835	0.363	0.000
98	2.018	0.400	0.199
94	0.605	0.443	0.000
110	2.020	0.552	0.000
101	1.697	0.625	0.213
95	3.039	0.879	0.000
103	1.195	0.897	0.247
105	2.430	1.015	0.177
104	2.479	1.083	0.000
114	2.412	1.104	0.000
93	2.431	1.187	0.000
72	2.539	1.188	0.000
111	2.597	1.272	0.000
71	2.044	1.273	0.000
112	2.881	1.445	0.000
75	2.355	1.530	0.000
74	2.368	1.559	0.000
119	2.038	1.606	0.000
124	1.982	1.741	0.000
117	0.779	1.811	0.000
77	1.607	1.938	0.134
79	1.253	2.134	0.000
133	1.241	2.226	0.000
131	1.770	2.292	0.000
130	2.102	2.533	0.177
123	1.245	2.575	0.159
132	2.022	2.594	0.000
137	2.024	2.898	0.000
136	2.283	3.400	0.000

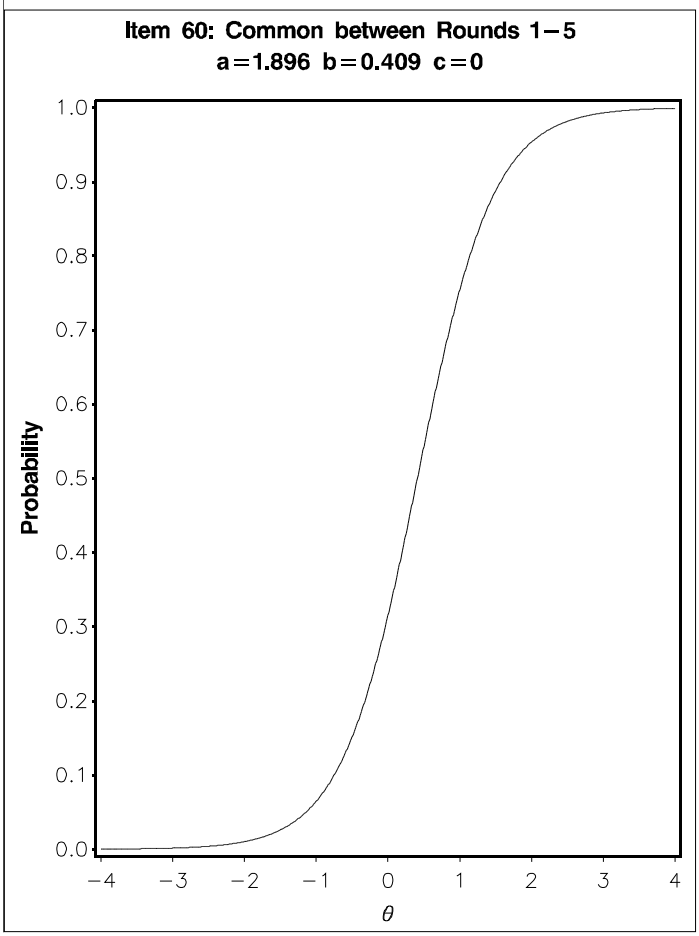
Appendix D

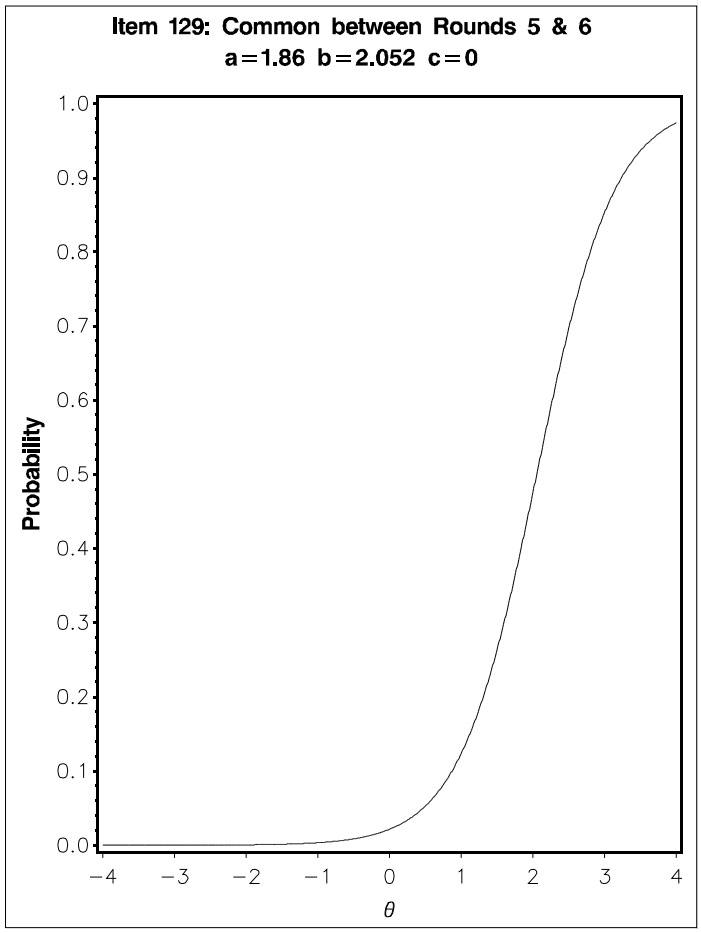
Unique Items in Round 6 Testing

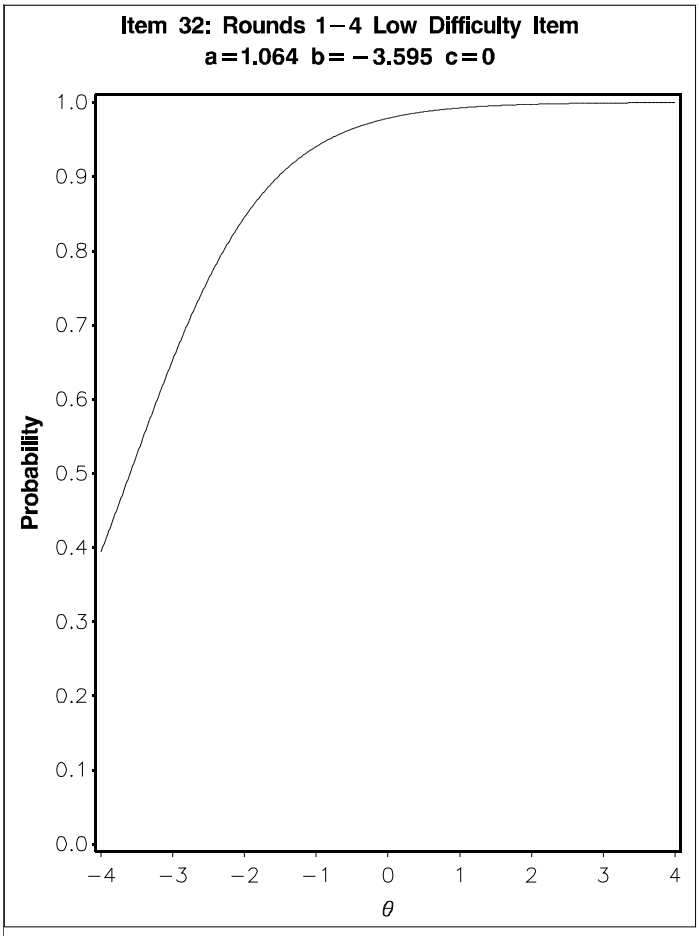
Items	a	b	c
182	1.4310	1.4130	0.0000
203	1.3240	1.4660	0.0000
197	2.3200	1.5030	0.0000
196	2.6810	1.6130	0.0000
204	2.0830	1.7150	0.0000
193	1.6050	1.7610	0.2057
205	1.1500	1.8310	0.0000
201	1.5610	1.8890	0.0000
171	2.0190	2.1110	0.0000
177	2.5030	2.4190	0.2076
178	3.3130	2.4580	0.0000
209	1.6900	2.5820	0.0000
208	1.6020	2.7330	0.0000
206	1.2530	2.7970	0.2565
210	1.9730	2.9000	0.0000
212	1.9880	2.9000	0.0000
216	2.9980	3.0910	0.0000
180	2.5990	3.3070	0.1271
213	2.0320	3.3330	0.1545
219	2.6410	3.3810	0.1512
217	1.8740	3.3910	0.0000
218	2.4360	3.4230	0.1185
181	1.3050	3.4290	0.2216
214	2.3610	3.5790	0.1166
215	2.2020	3.6590	0.0963
226	2.7490	3.7090	0.0000
184	1.8040	3.9900	0.0000
221	1.5940	4.3320	0.0000
222	2.1170	4.3830	0.0000

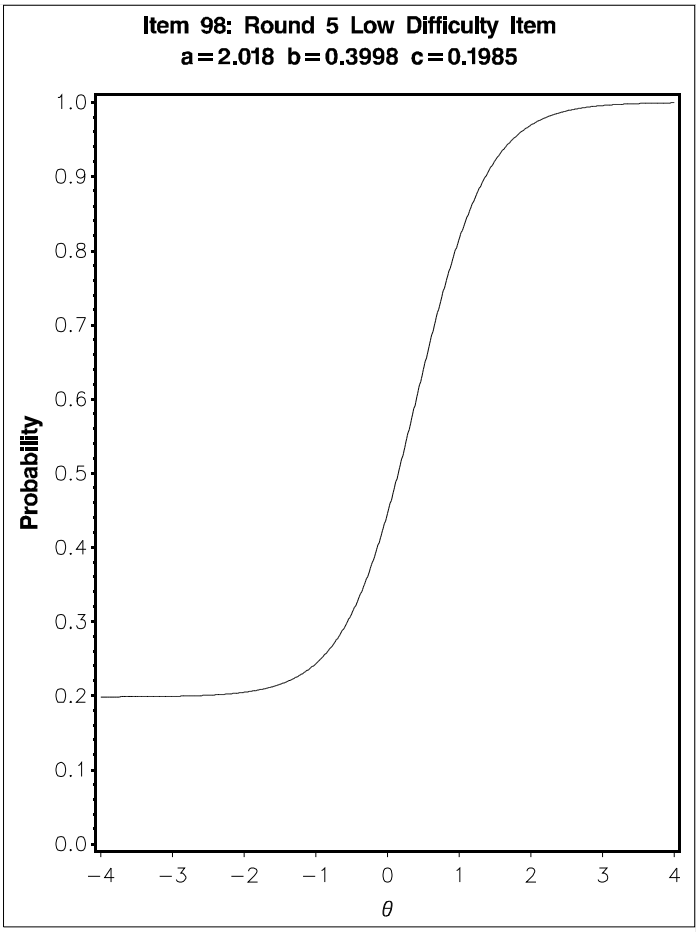

Item 13: Constrained Item
a=1.8 b=0.35 c=0
Probability
θ


Item 14: Common across ALL rounds
a=1.899 b=0.3571 c=0
Probability
θ


Item 60: Common between Rounds 1–5
a=1.896 b=0.409 c=0
Probability
θ


Item 129: Common between Rounds 5 & 6
a=1.86 b=2.052 c=0
Probability
θ


Item 32: Rounds 1−4 Low Difficulty Item
a=1.064 b=−3.595 c=0
Probability
θ


Item 98: Round 5 Low Difficulty Item
a=2.018 b=0.3998 c=0.1985
Probability
θ

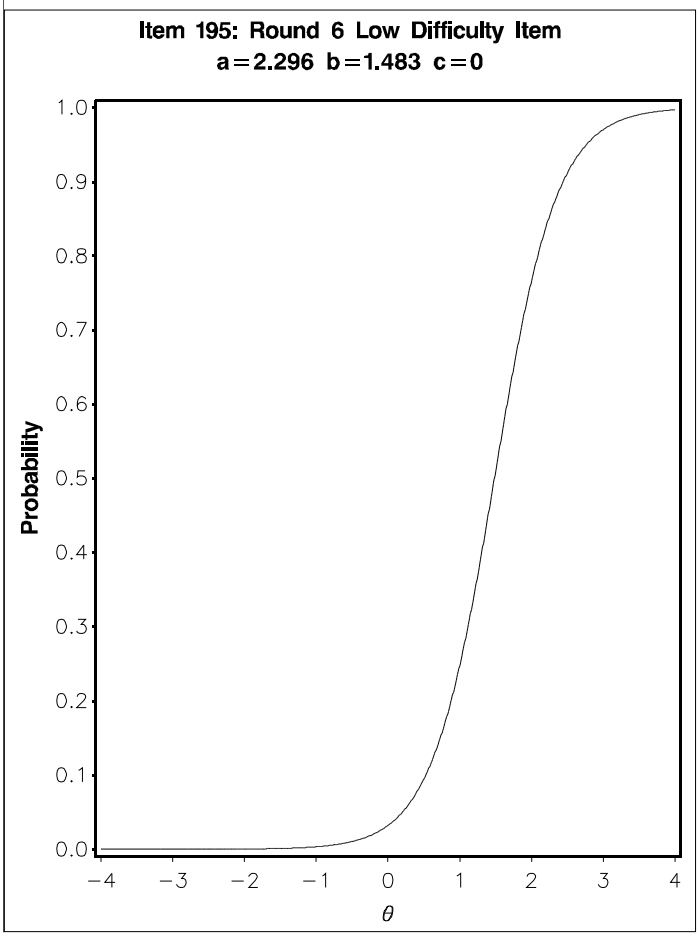
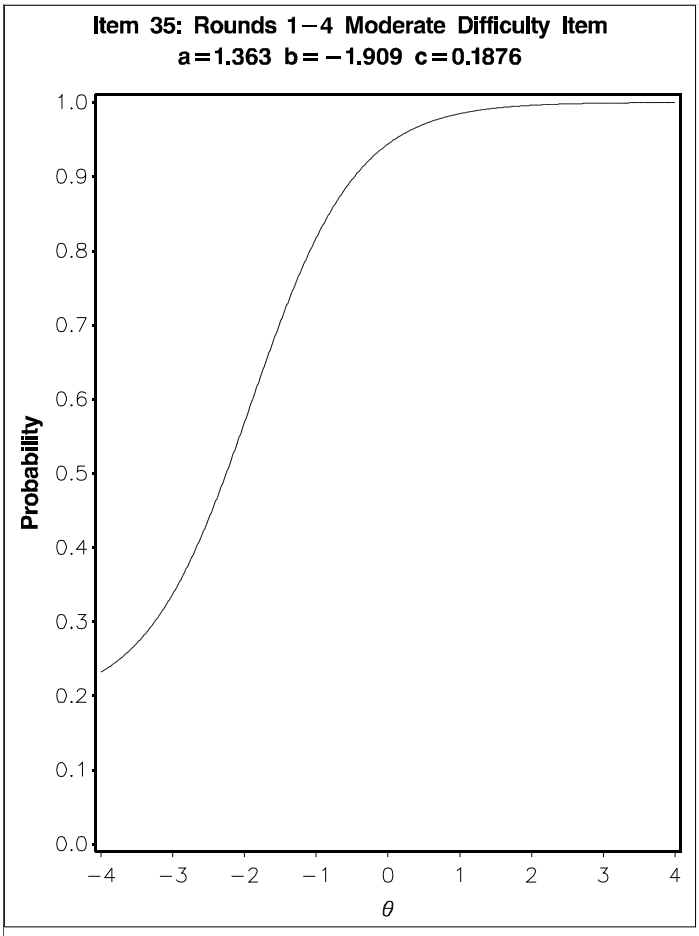
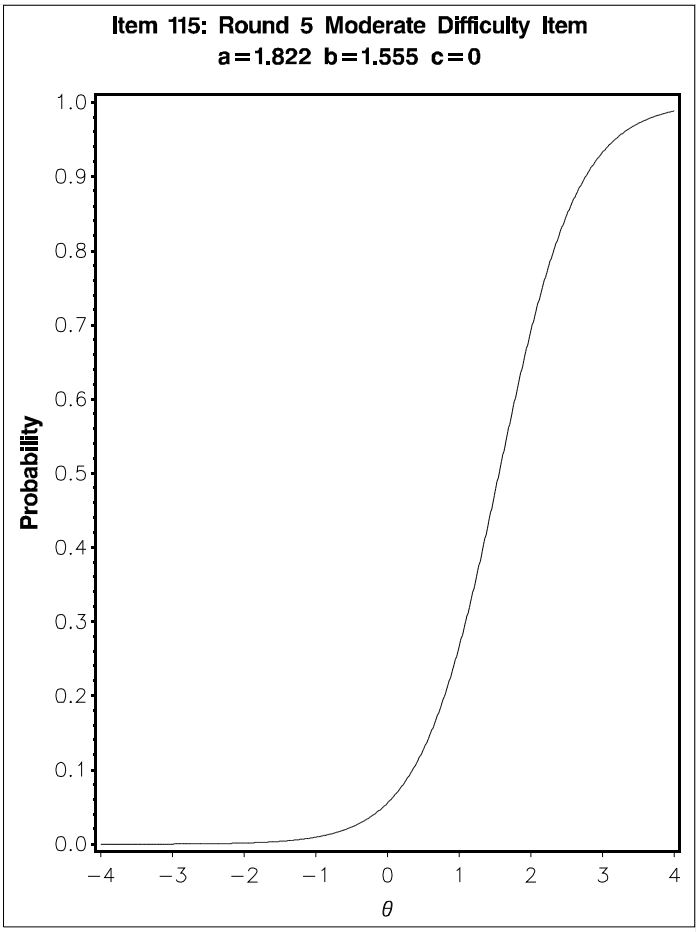

Item 195: Round 6 Low Difficulty Item
a=2.296 b=1.483 c=0
Probability
θ

Figure 3


Item 35: Rounds 1−4 Moderate Difficulty Item
a=1.363 b=−1.909 c=0.1876
Probability
θ


Item 115: Round 5 Moderate Difficulty Item
a=1.822 b=1.555 c=0
Probability
θ

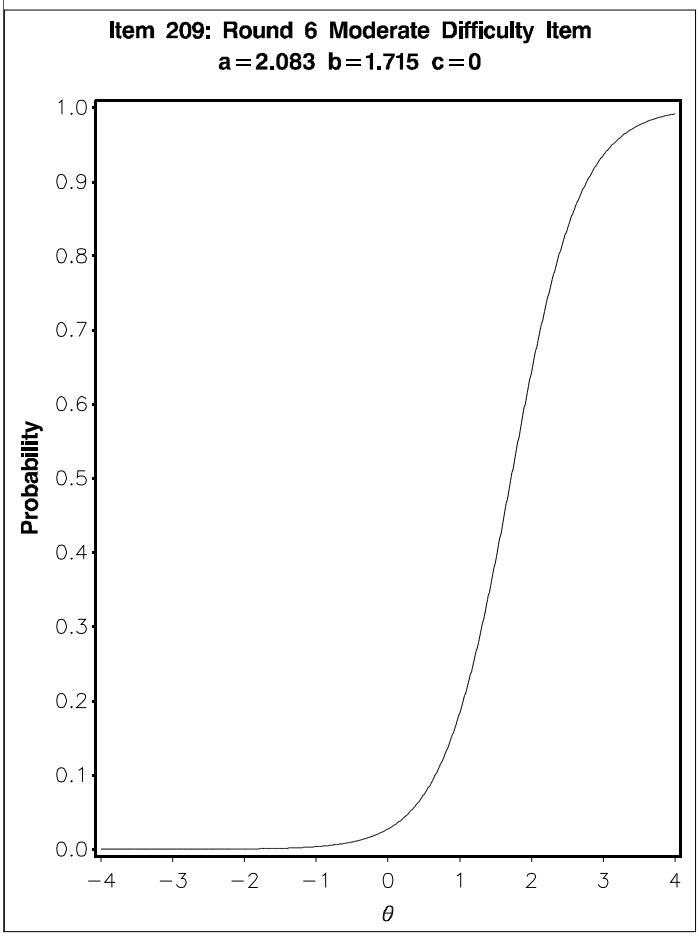
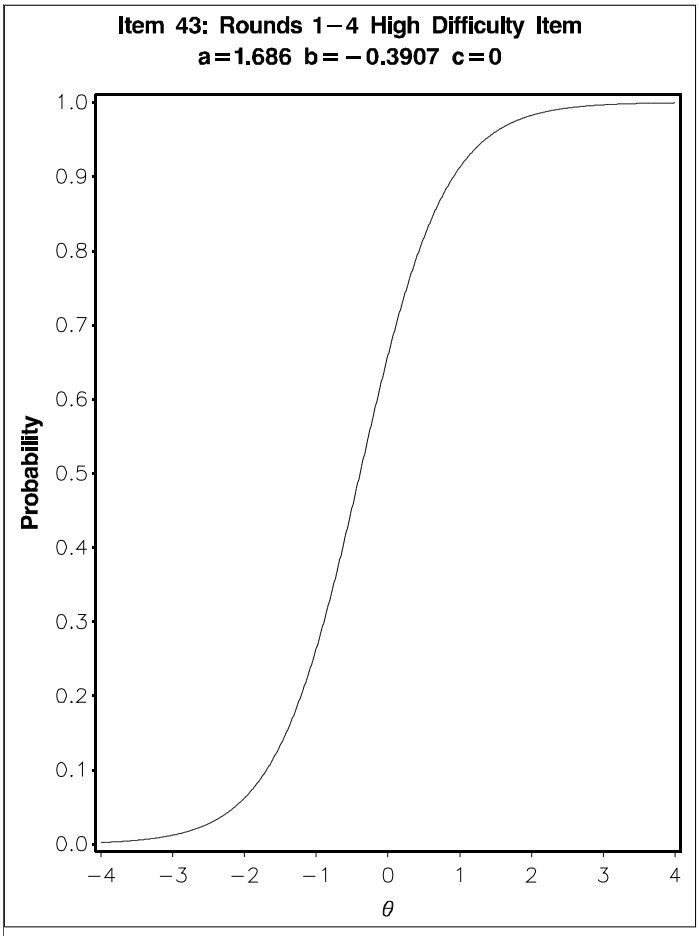
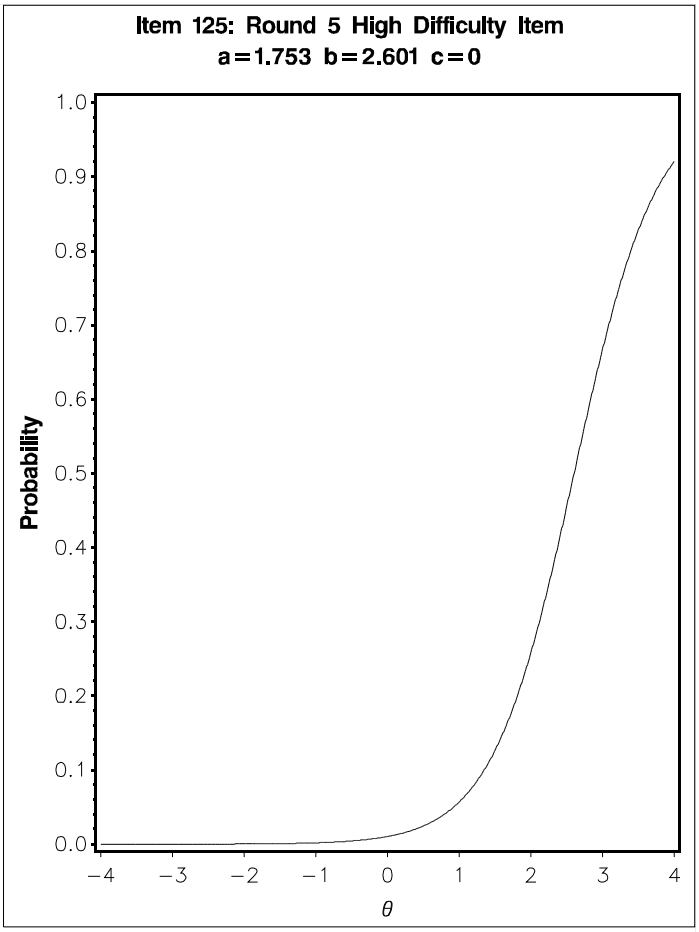

Item 209: Round 6 Moderate Difficulty Item
a=2.083 b=1.715 c=0
Probability
θ

Figure 4


Item 43: Rounds 1−4 High Difficulty Item
a=1.686 b=−0.3907 c=0
Probability
θ


Item 125: Round 5 High Difficulty Item
a=1.753 b=2.601 c=0
Probability
θ

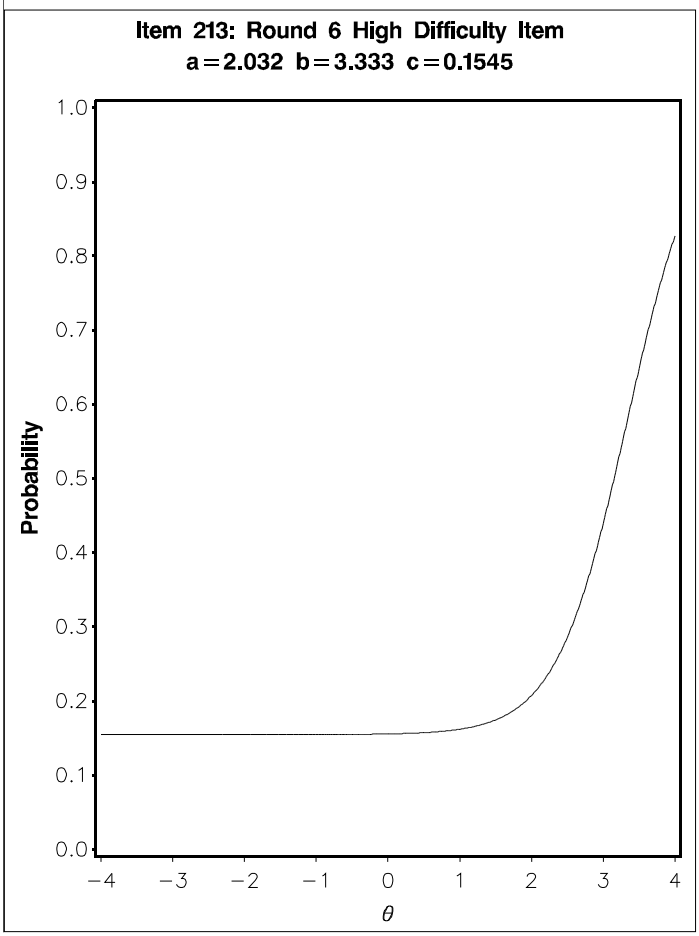
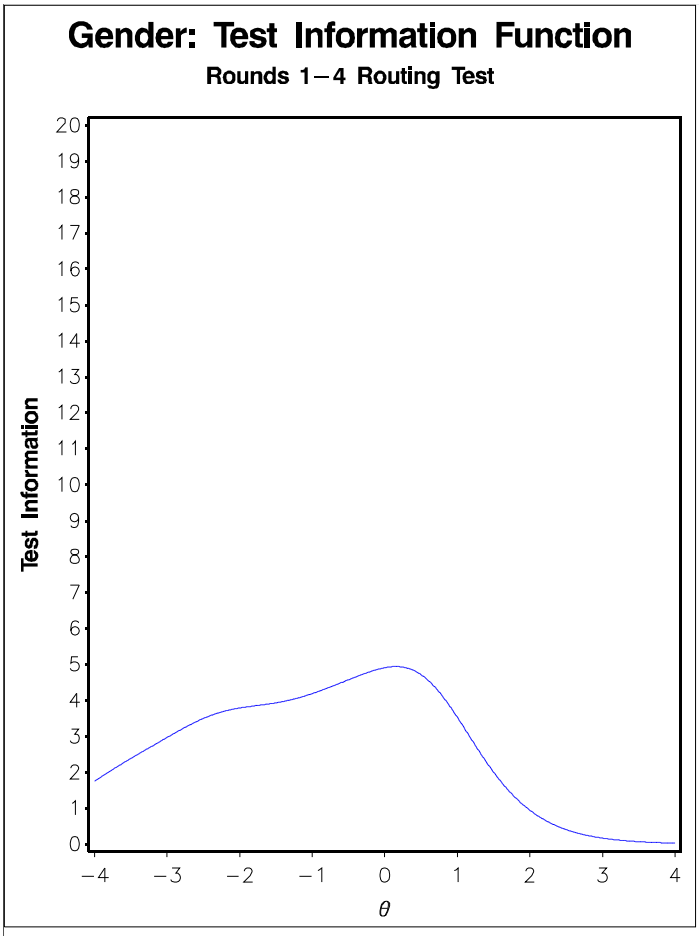

Item 213: Round 6 High Difficulty Item
a=2.032 b=3.333 c=0.1545
Probability
θ

Figure 5



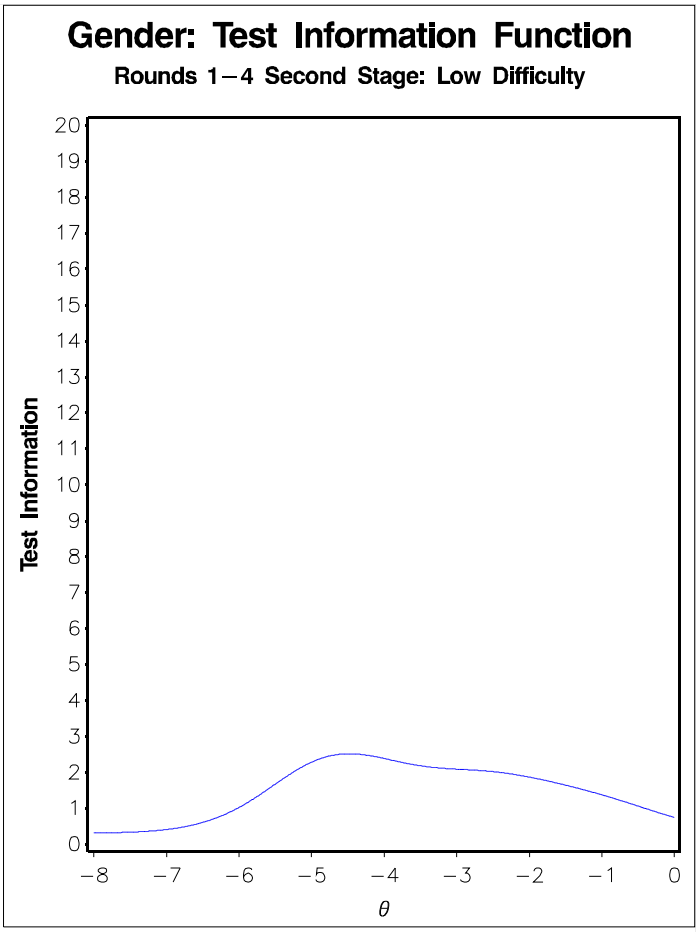
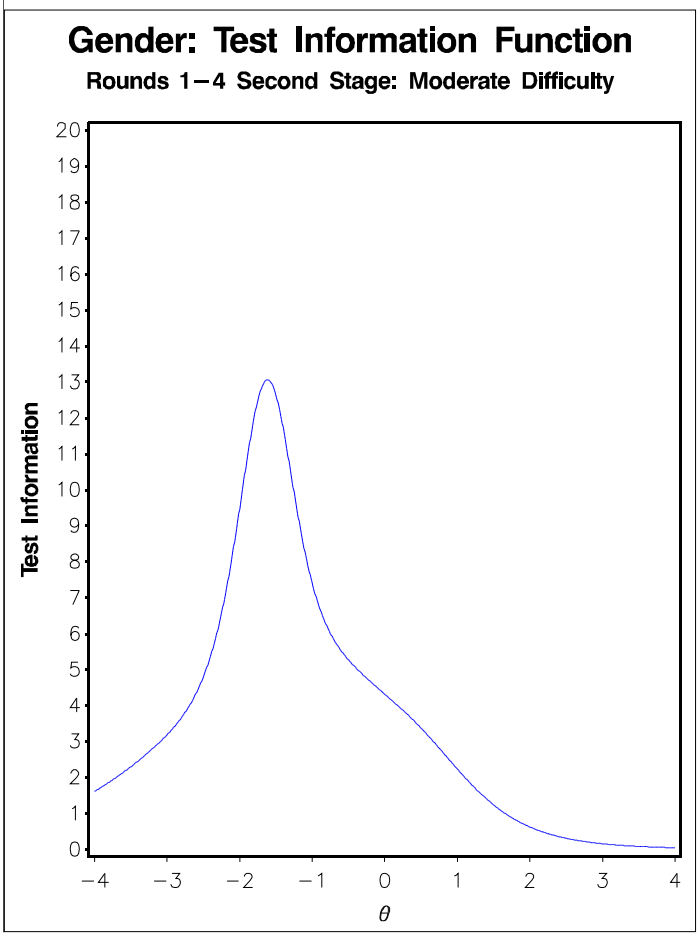
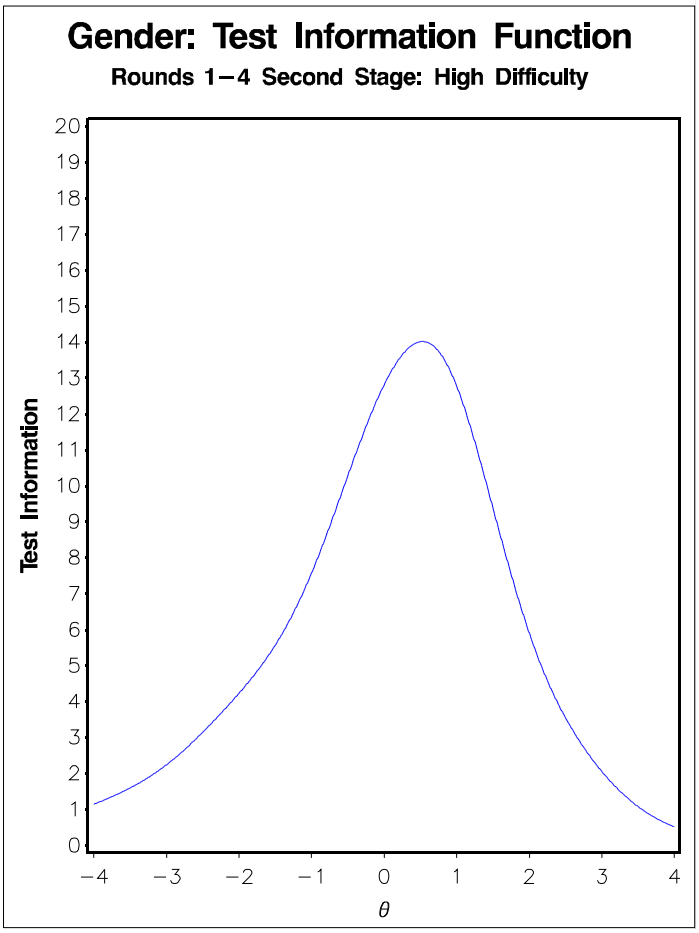
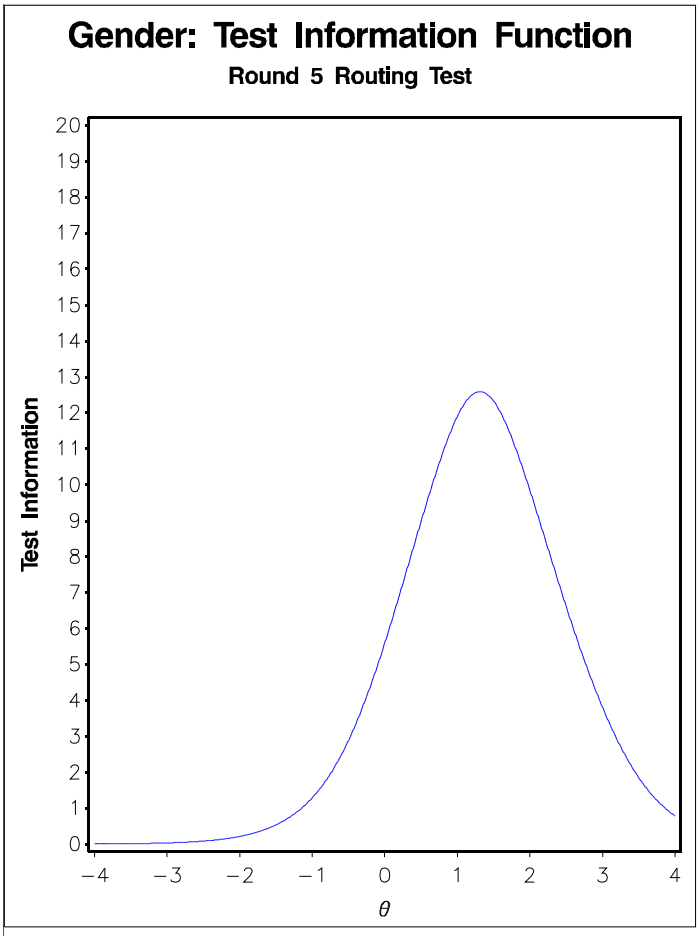
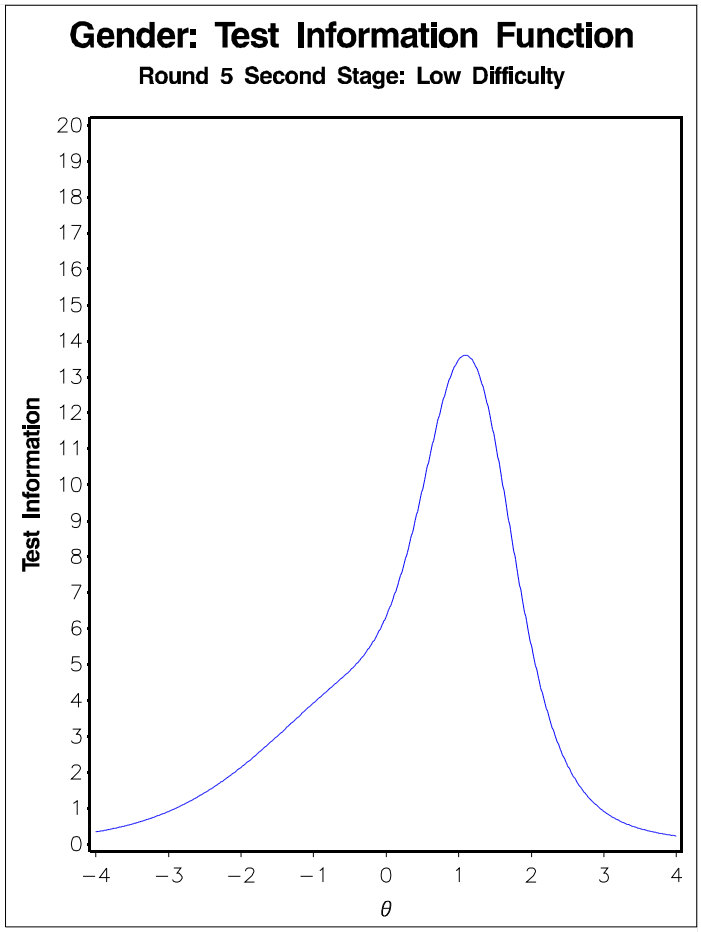


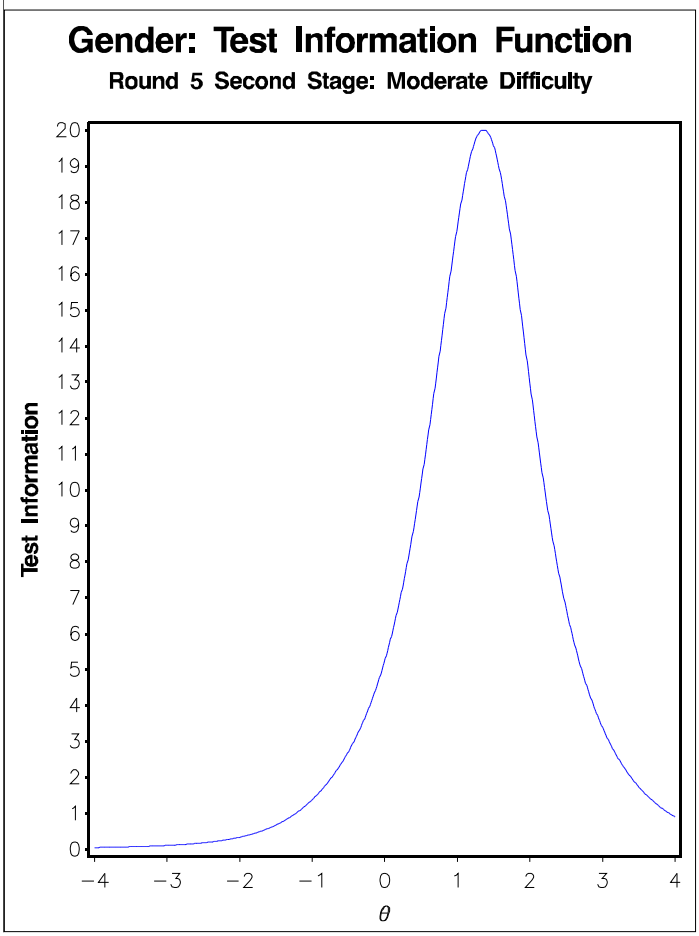
Figure 3






Gender: Test Information Function
Round 5 Routing Test
Test Information
θ


Gender: Test Information Function
Round 5 Second Stage: Low Difficulty
Test Information
θ


Gender: Test Information Function
Round 5 Second Stage: Moderate Difficulty
Test Information
θ

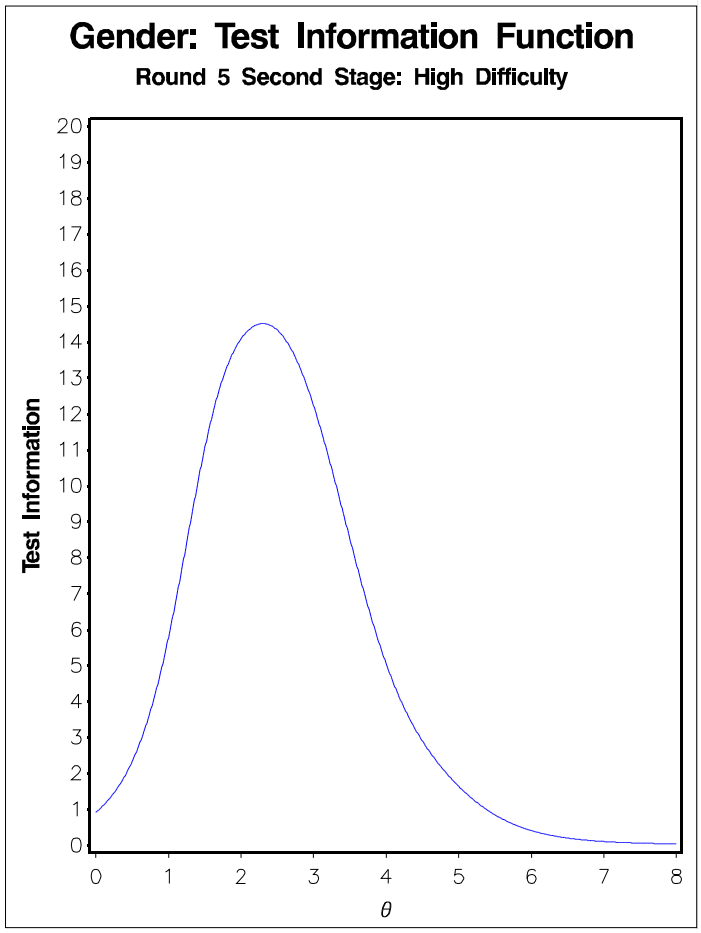
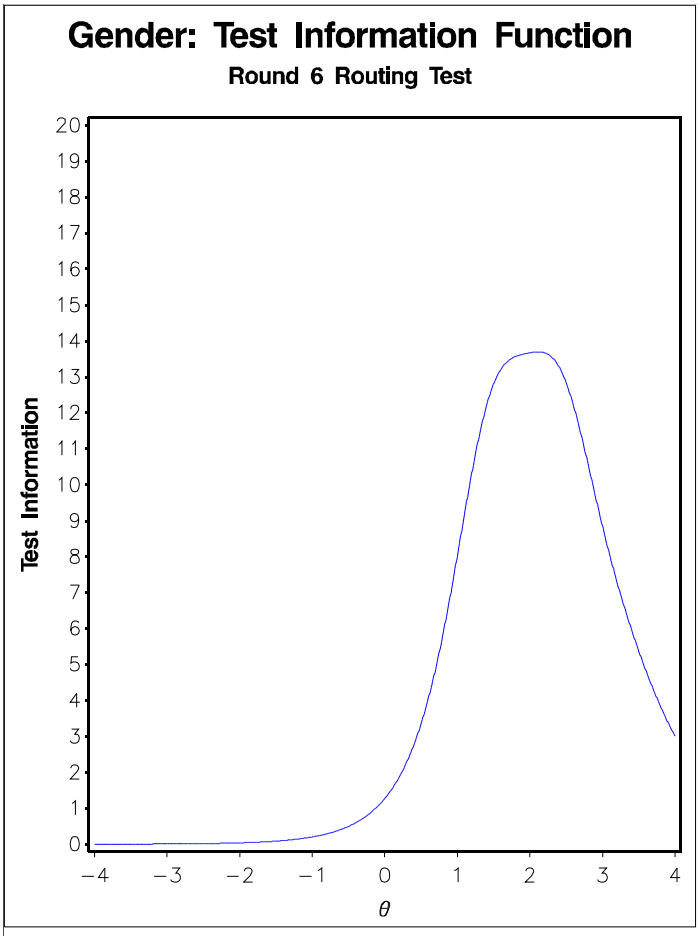
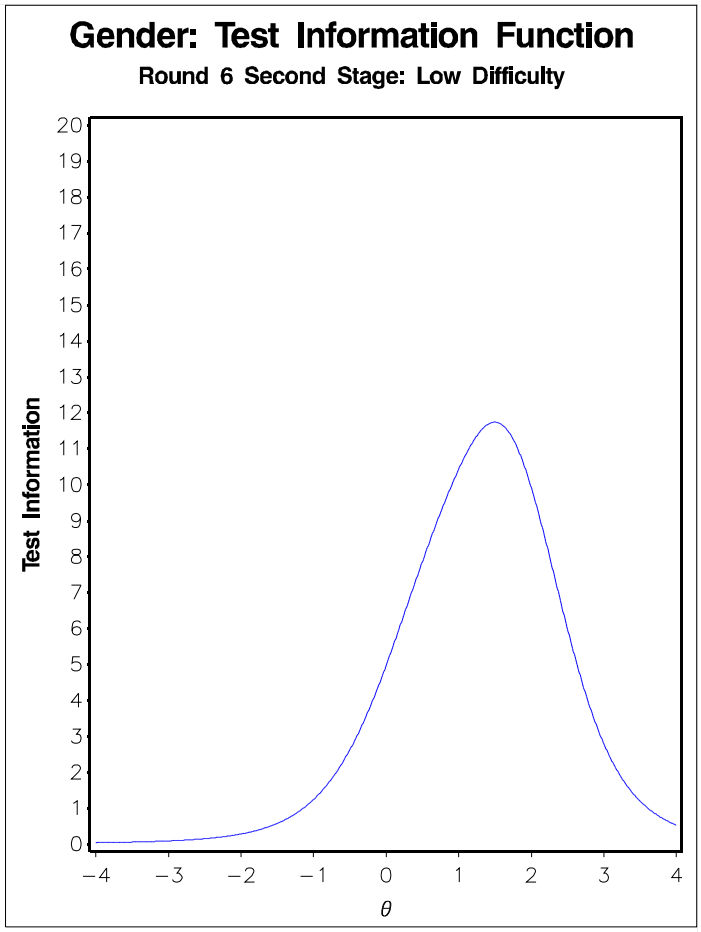
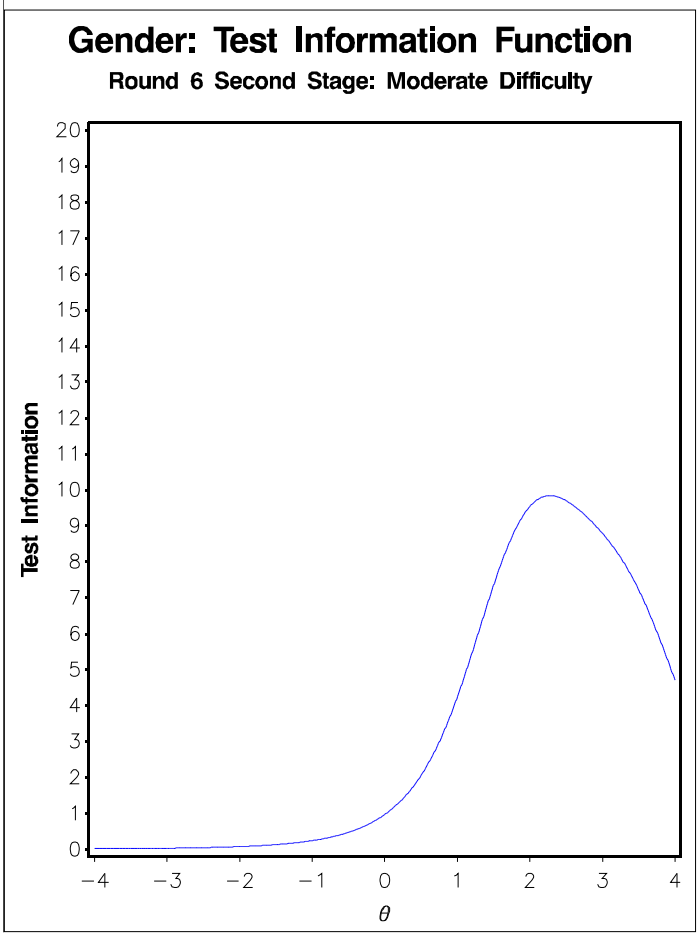
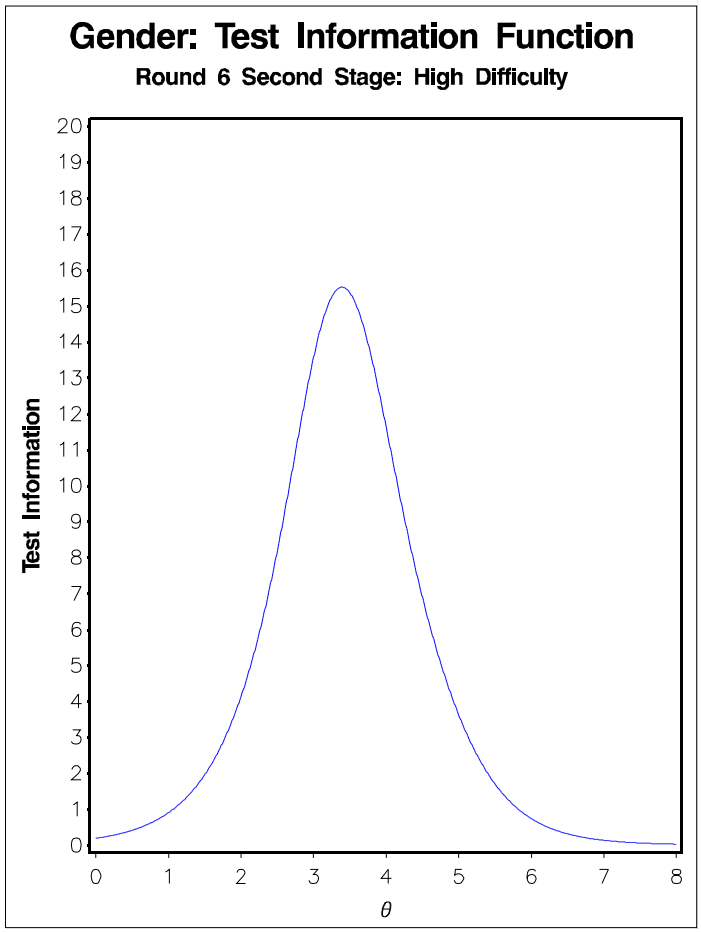

Gender: Test Information Function
Round 5 Second Stage: High Difficulty
Test Information
θ

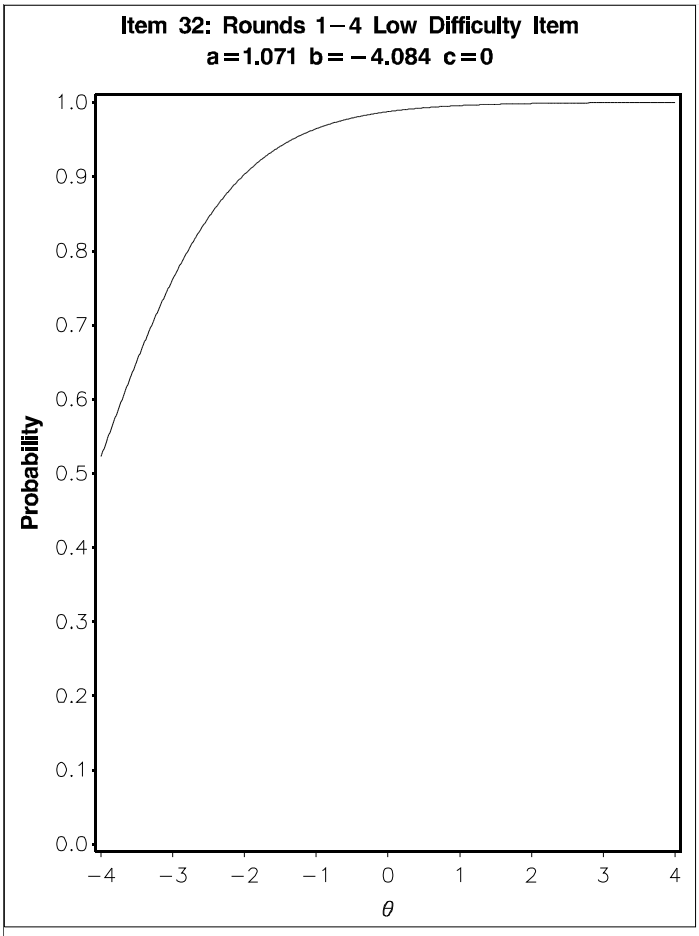
Figure 7

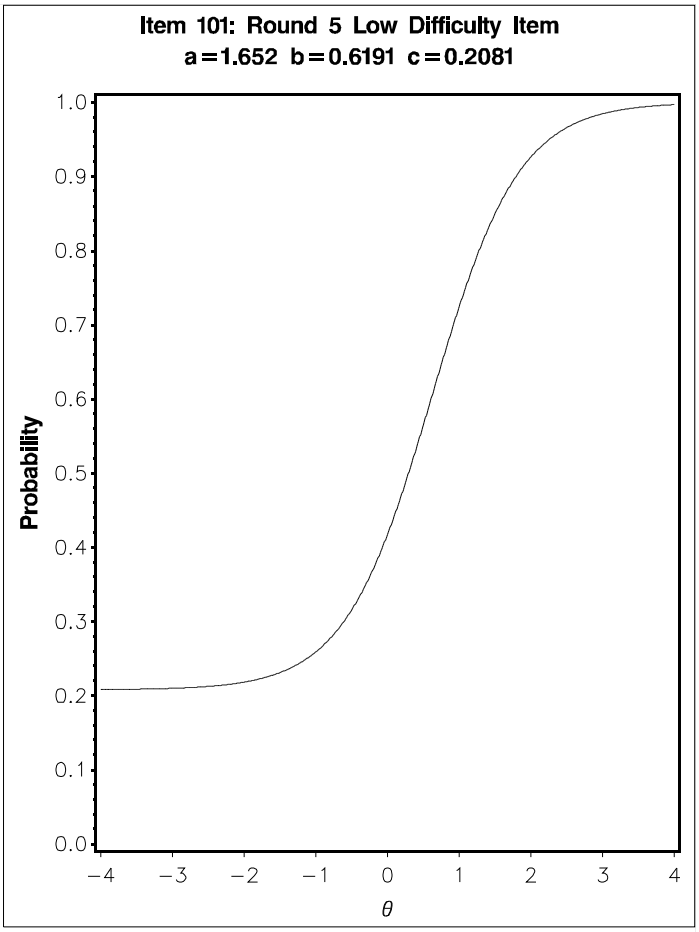

Gender: Test Information Function
Round 6 Routing Test
Test Information
θ


Gender: Test Information Function
Round 6 Second Stage: Low Difficulty
Test Information
θ


Gender: Test Information Function
Round 6 Second Stage: Moderate Difficulty
Test Information
θ


Gender: Test Information Function
Round 6 Second Stage: High Difficulty
Test Information
θ


Item 32: Rounds 1−4 Low Difficulty Item
a=1.071 b=−4.084 c=0
Probability
θ


Item 101: Round 5 Low Difficulty Item
a=1.652 b=0.6191 c=0.2081
Probability
θ

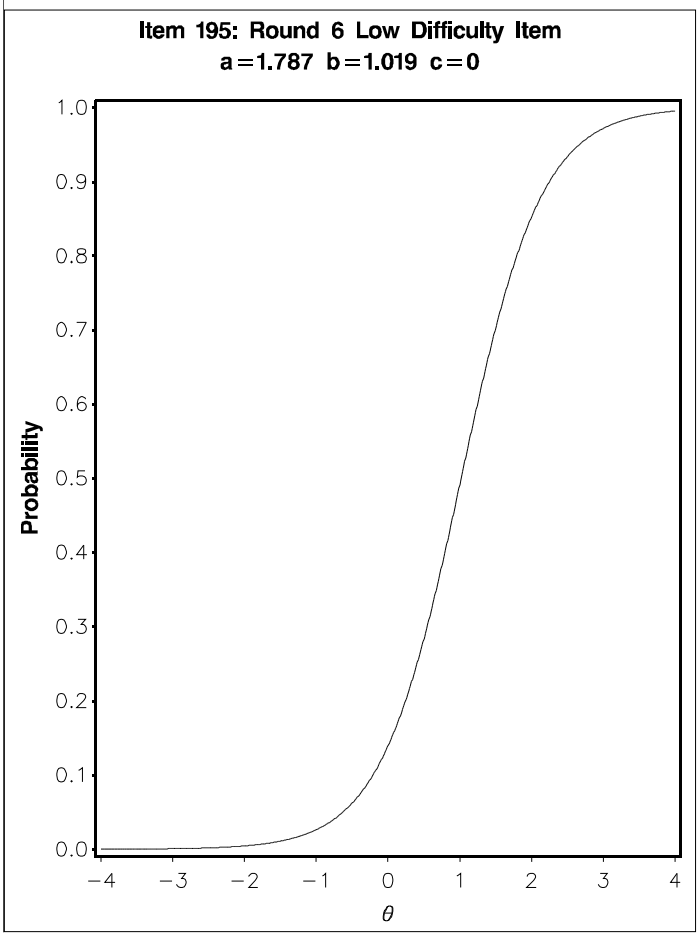
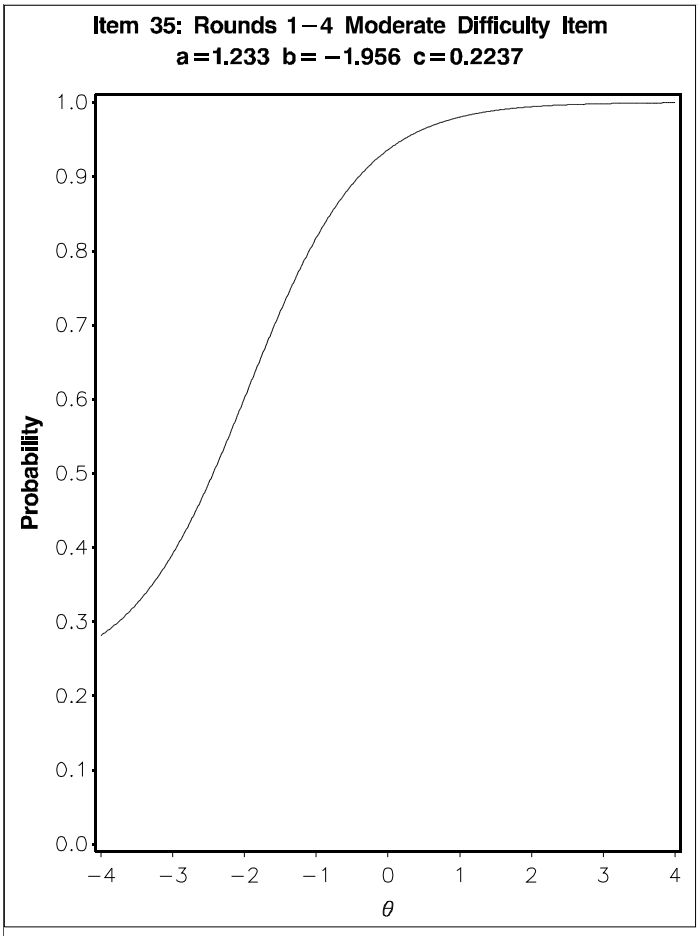
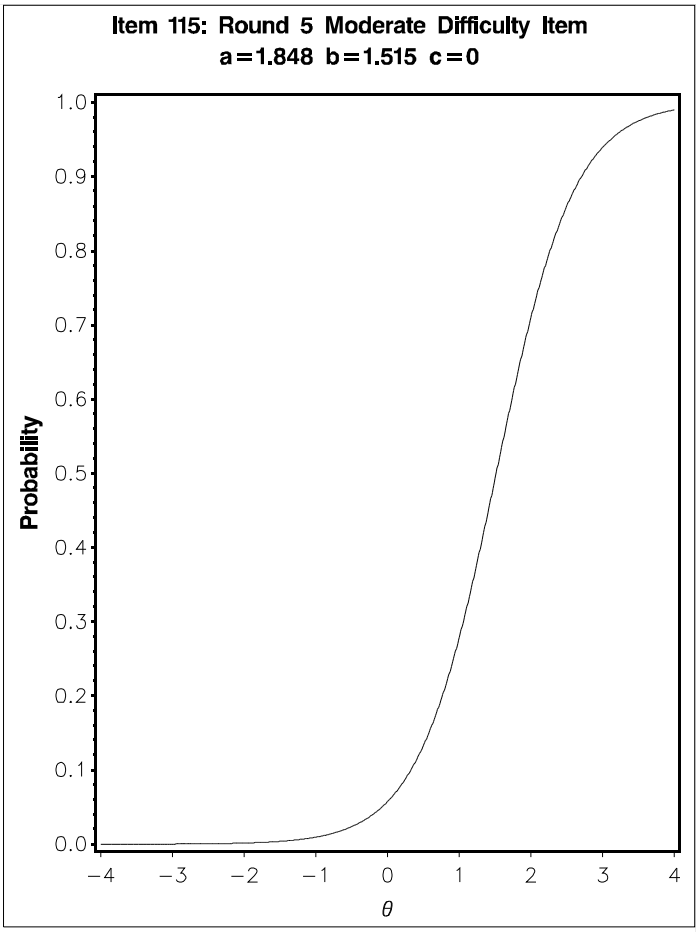

Item 195: Round 6 Low Difficulty Item
a=1.787 b=1.019 c=0
Probability
θ

Figure 9


Item 35: Rounds 1−4 Moderate Difficulty Item
a=1.233 b=−1.956 c=0.2237
Probability
θ


Item 115: Round 5 Moderate Difficulty Item
a=1.848 b=1.515 c=0
Probability
θ

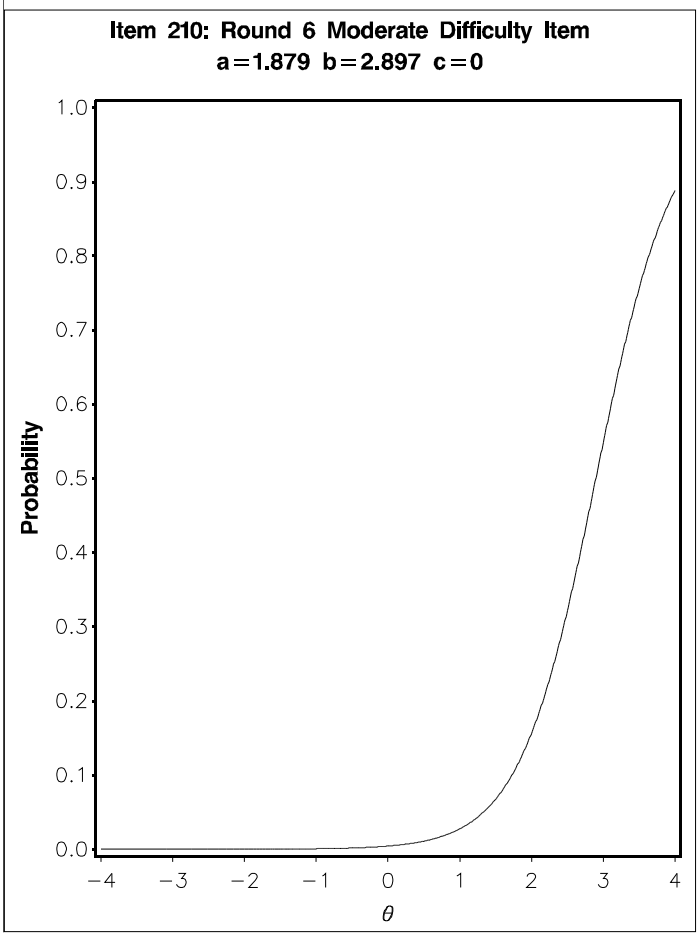
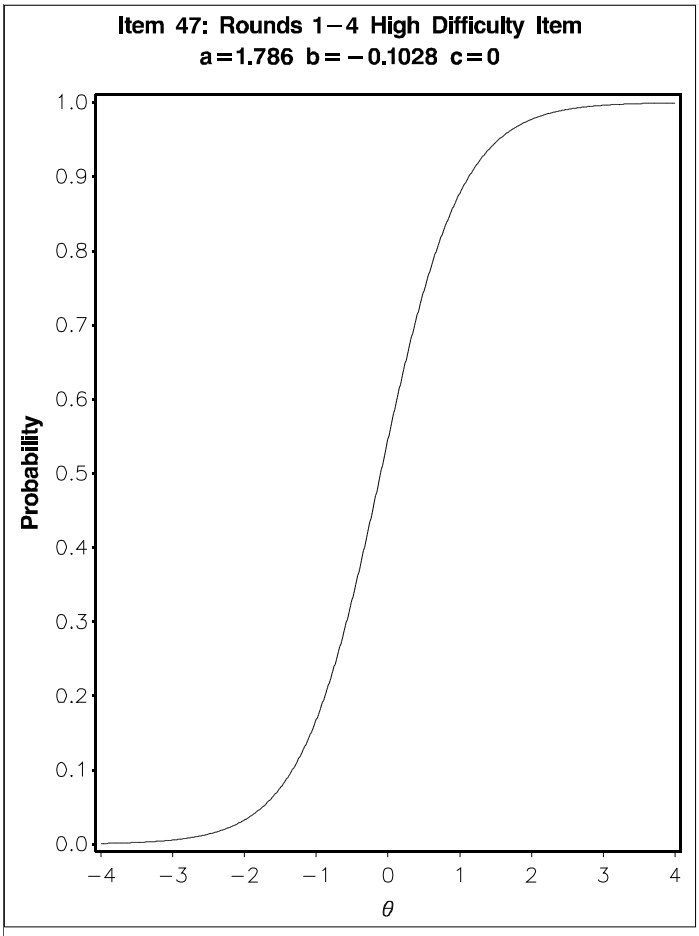
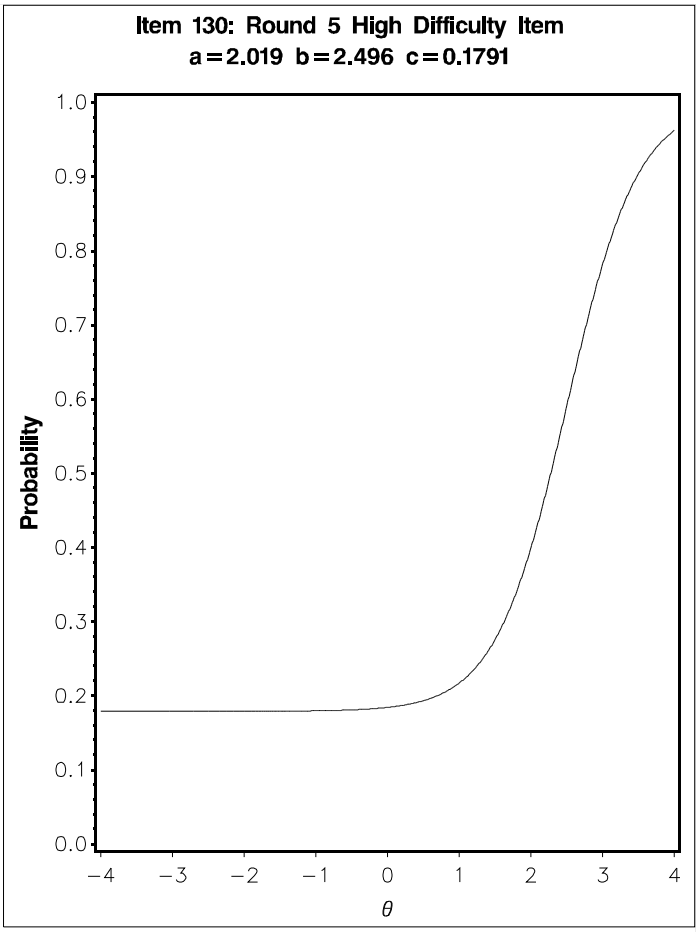
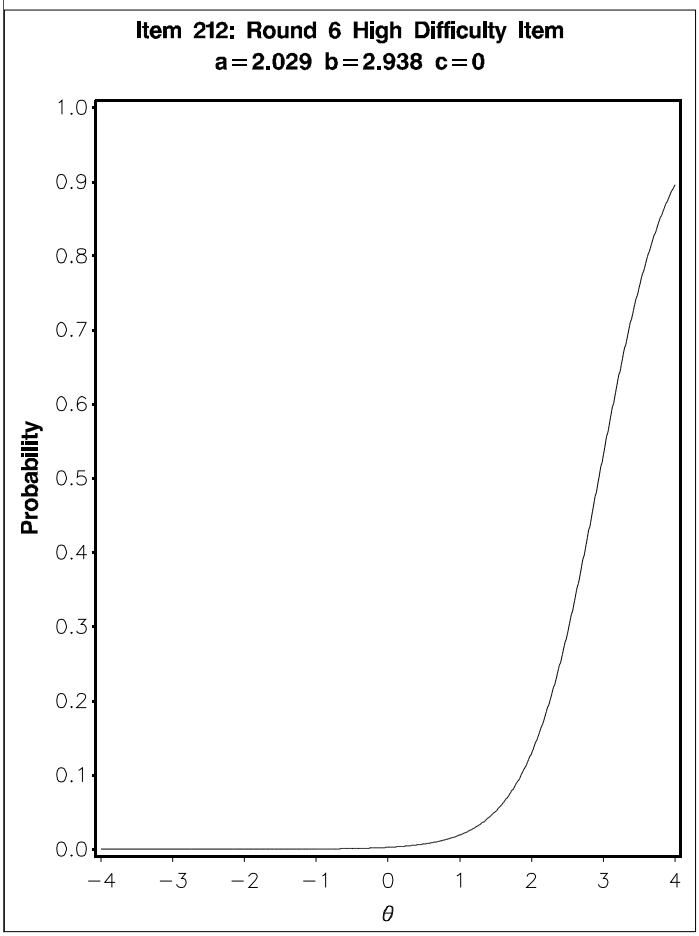
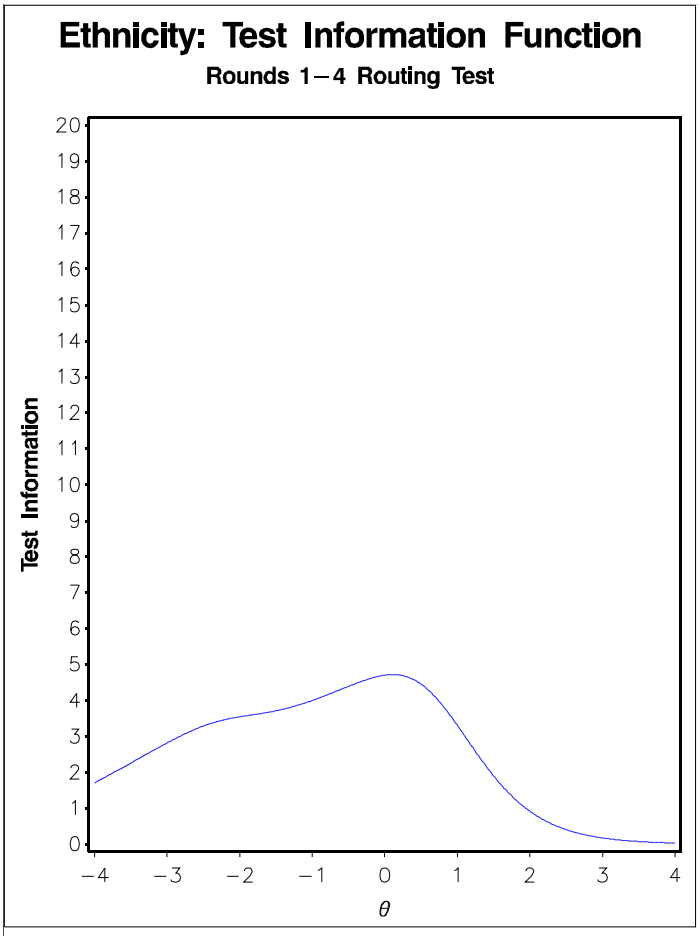

Item 210: Round 6 Moderate Difficulty Item
a=1.879 b=2.897 c=0
Probability
θ

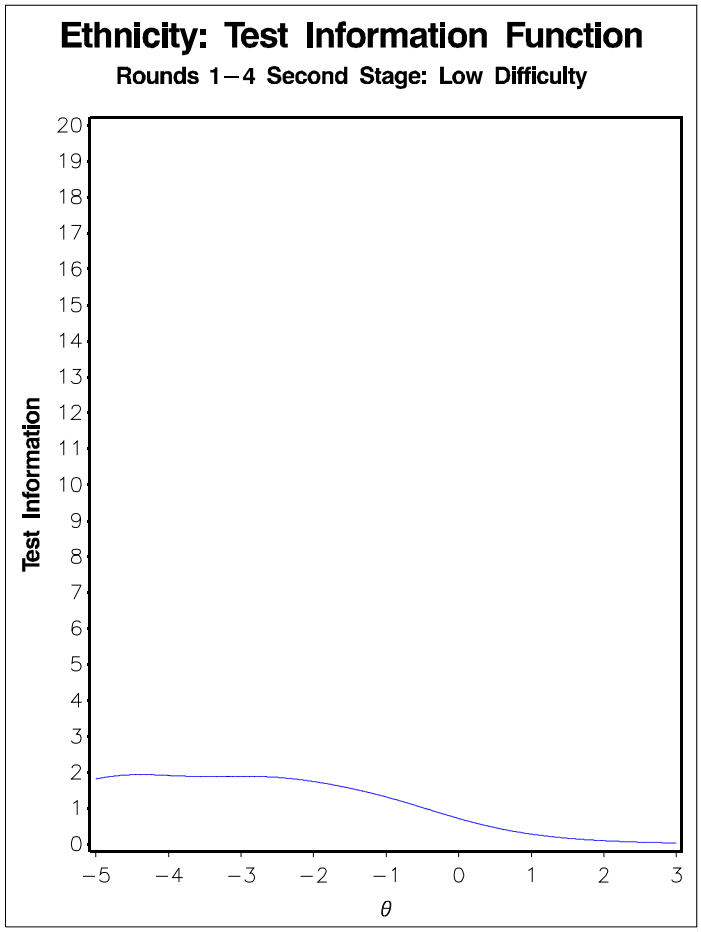
Figure 10

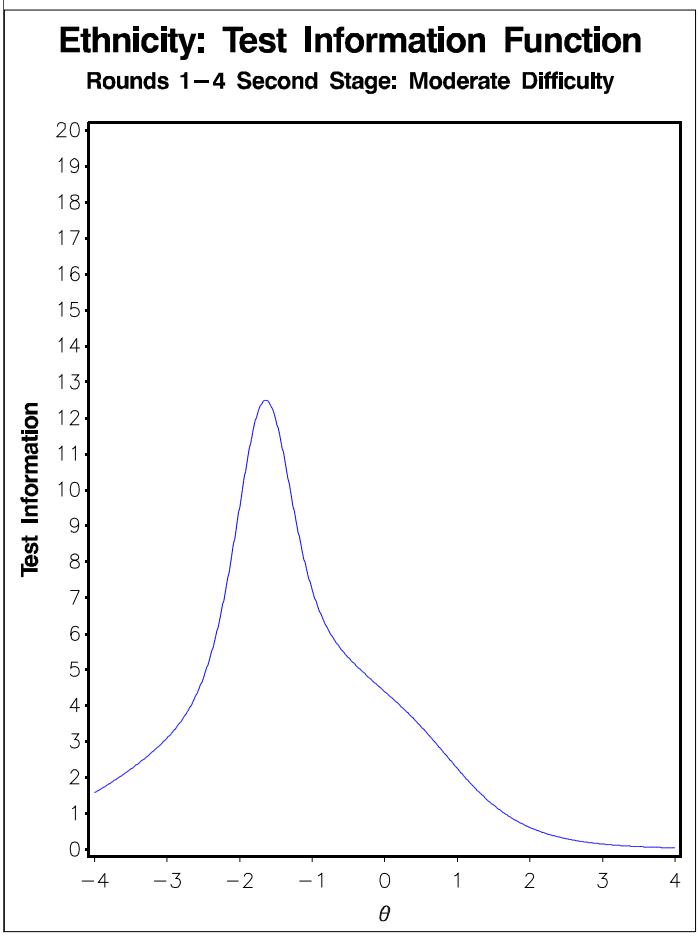

Item 47: Rounds 1−4 High Difficulty Item
a=1.786 b=−0.1028 c=0
Probability
θ

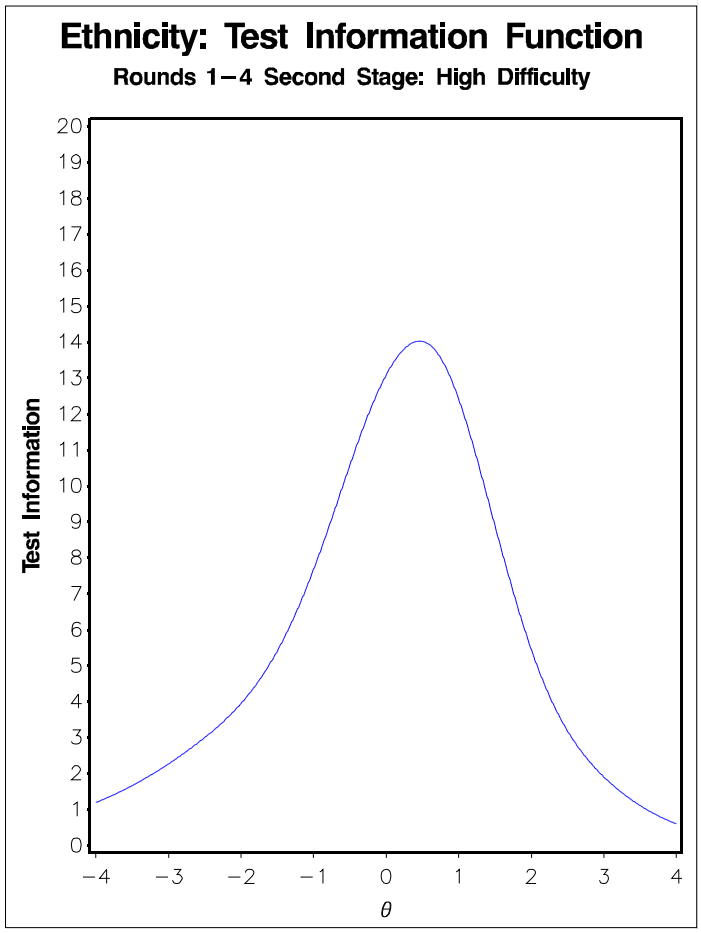

Item 130: Round 5 High Difficulty Item
a=2.019 b=2.496 c=0.1791
Probability
θ

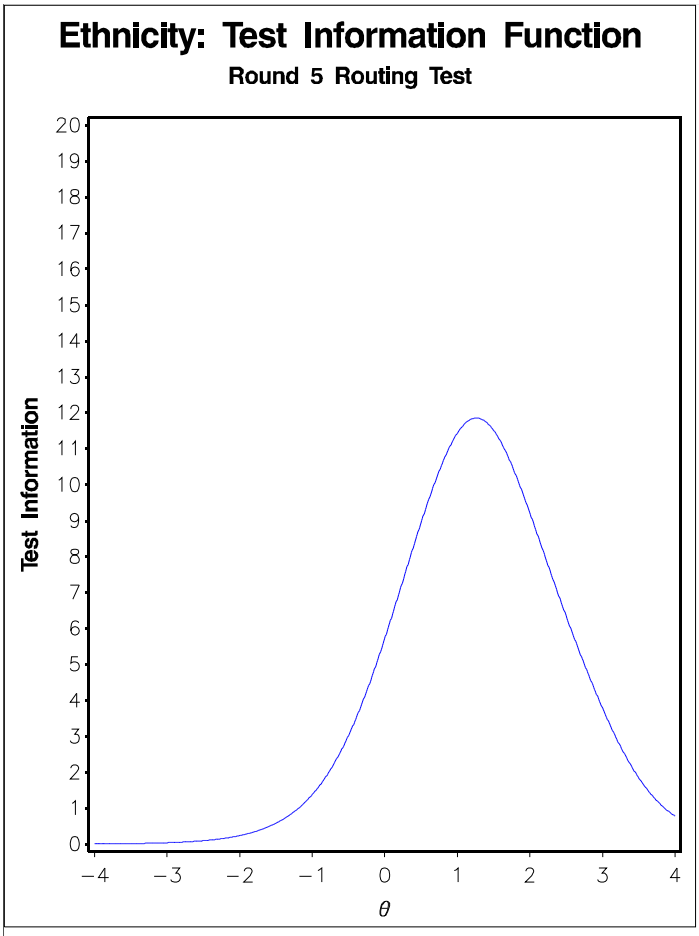

Item 212: Round 6 High Difficulty Item
a=2.029 b=2.938 c=0
Probability
θ

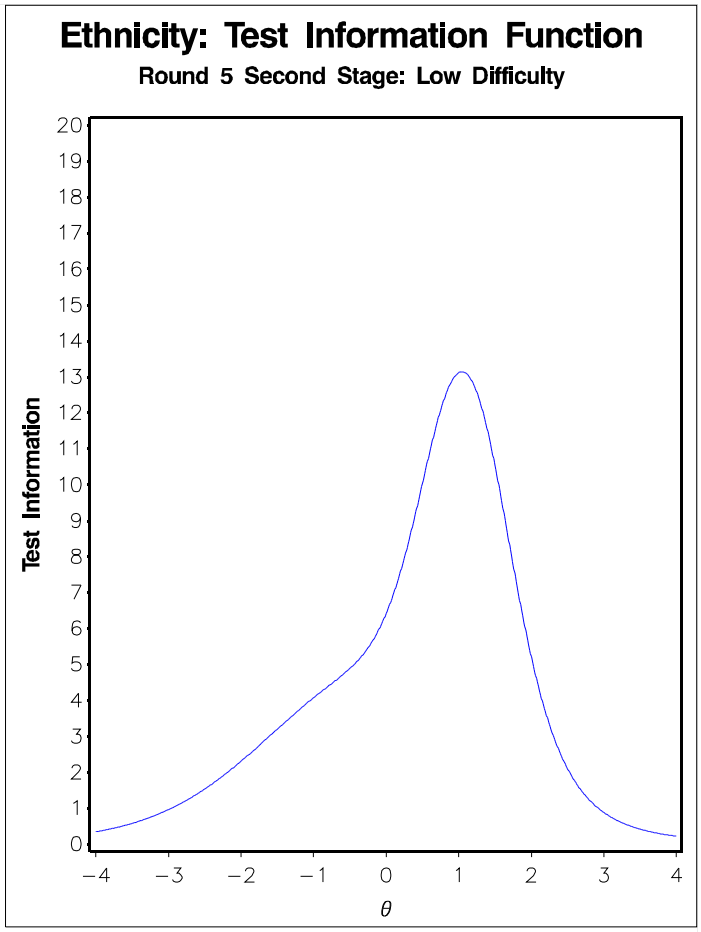

Ethnicity: Test Information Function
Rounds 1−4 Routing Test
Test Information
θ

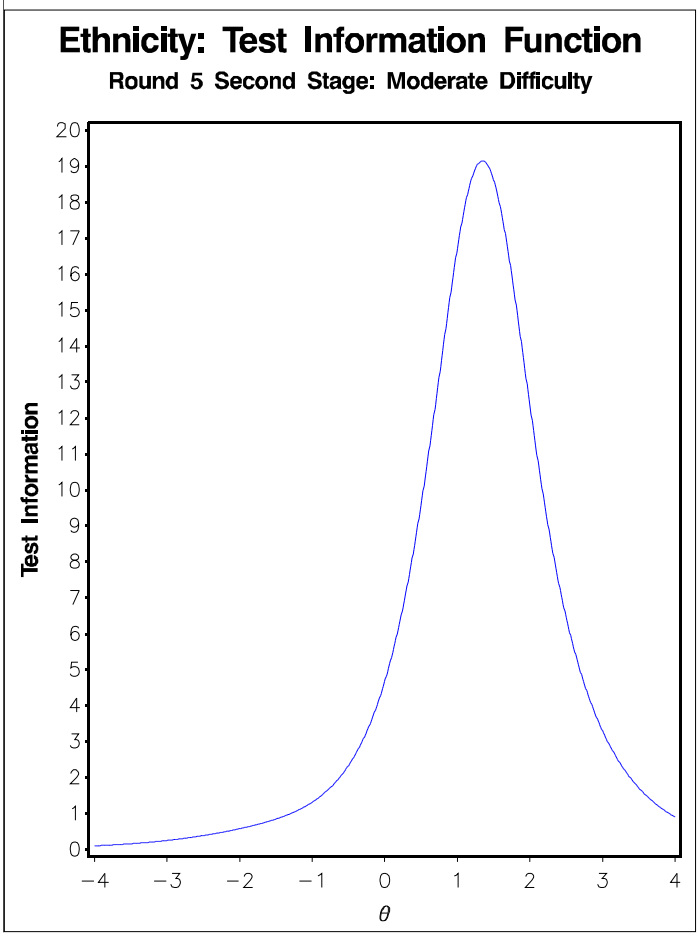

Ethnicity: Test Information Function
Rounds 1−4 Second Stage: Low Difficulty
Test Information
θ

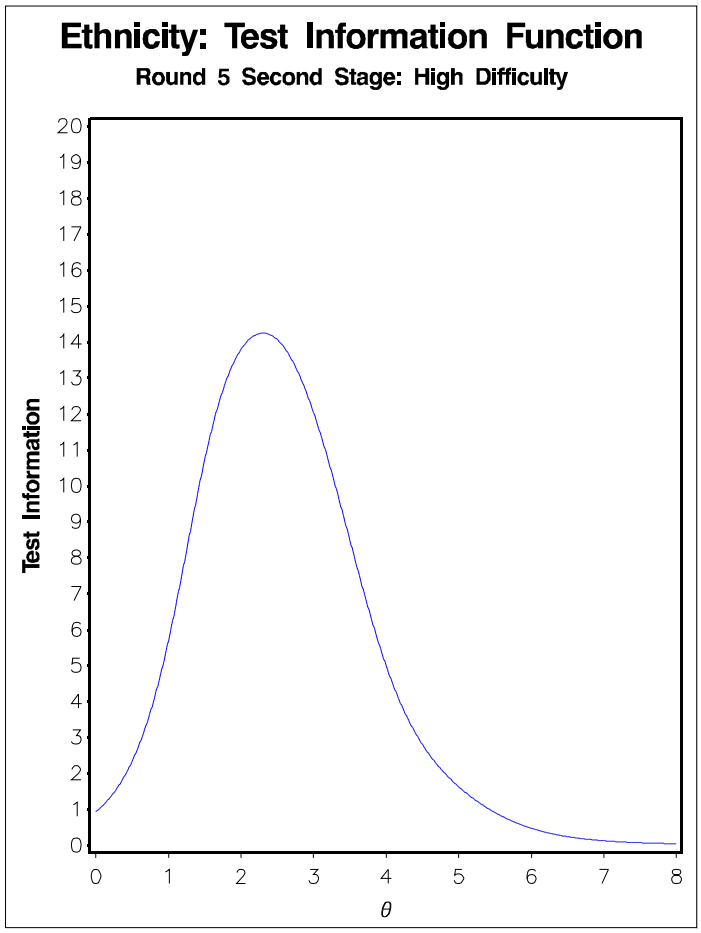

Ethnicity: Test Information Function
Rounds 1−4 Second Stage: Moderate Difficulty
Test Information
θ

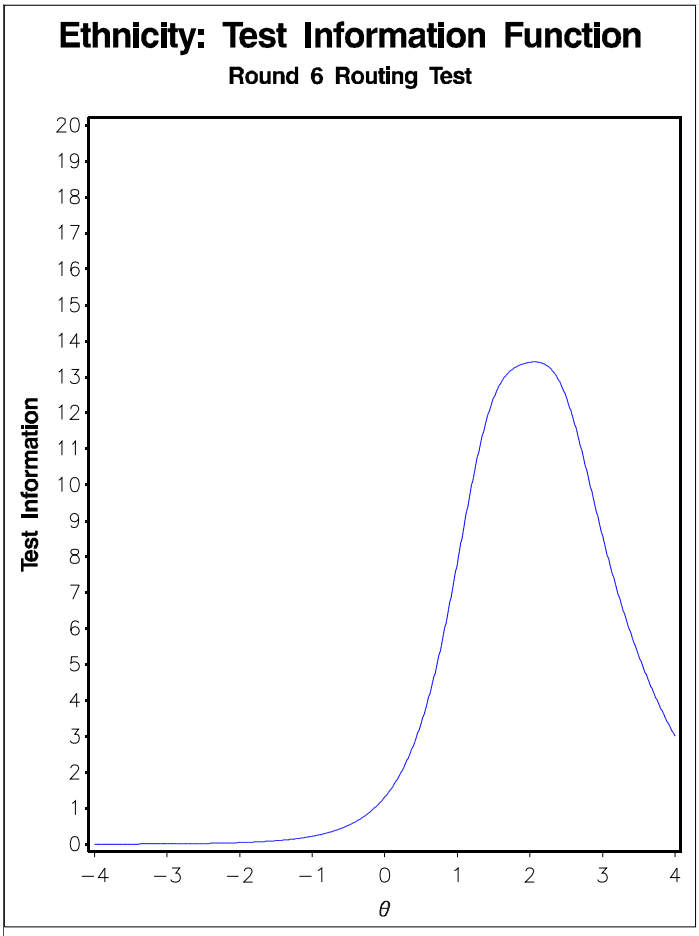

Ethnicity: Test Information Function
Rounds 1−4 Second Stage: High Difficulty
Test Information
θ

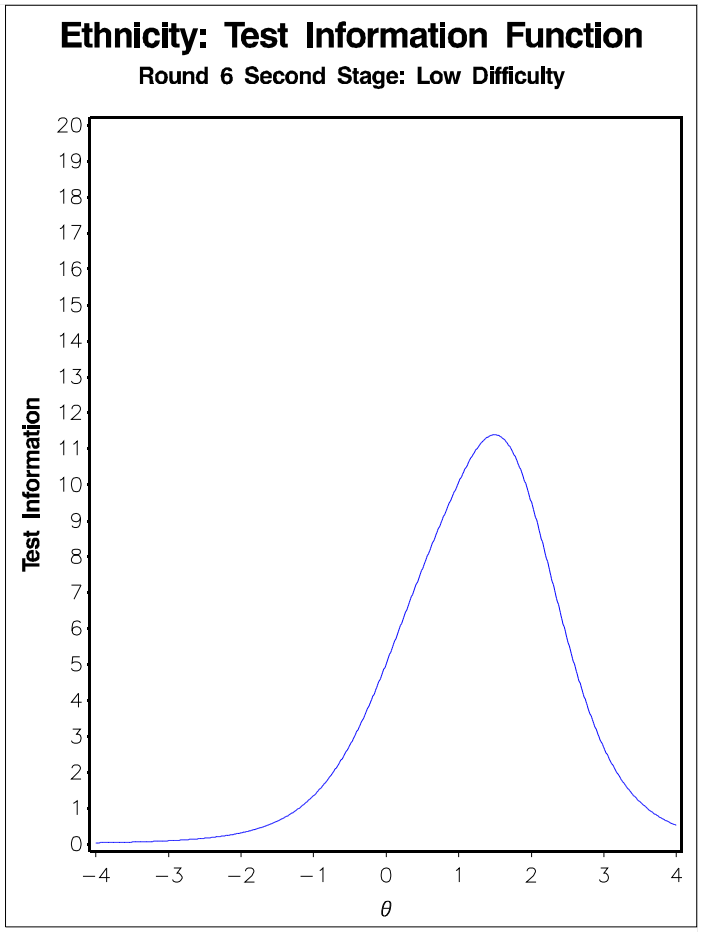

Ethnicity: Test Information Function
Round 5 Routing Test
Test Information
θ

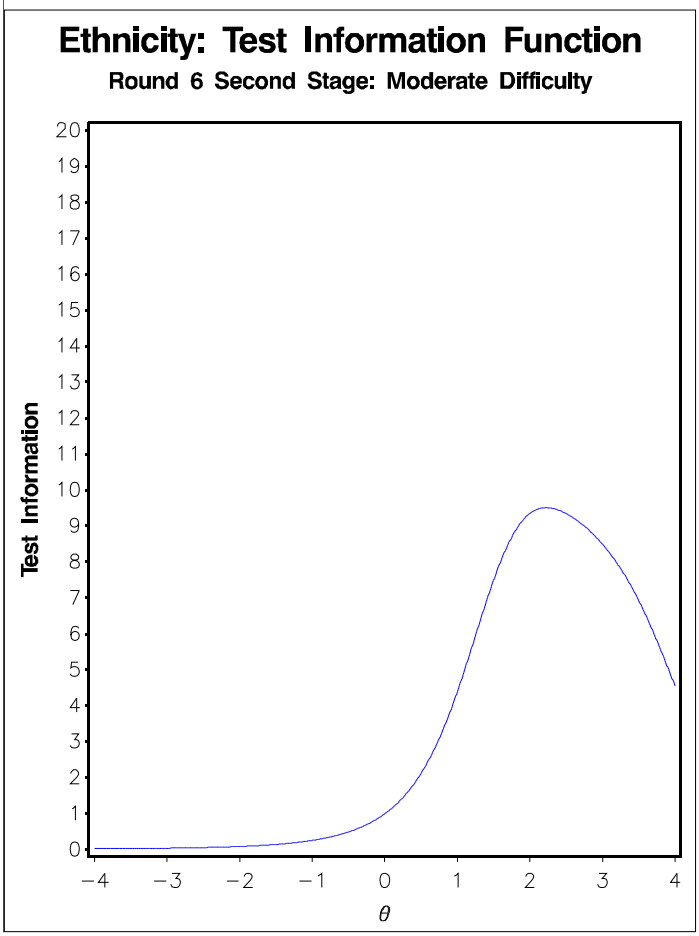

Ethnicity: Test Information Function
Round 5 Second Stage: Low Difficulty
Test Information
θ


Ethnicity: Test Information Function
Round 5 Second Stage: Moderate Difficulty
Test Information
θ


Ethnicity: Test Information Function
Round 5 Second Stage: High Difficulty
Test Information
θ


Ethnicity: Test Information Function
Round 6 Routing Test
Test Information
θ


Ethnicity: Test Information Function
Round 6 Second Stage: Low Difficulty
Test Information
θ


Ethnicity: Test Information Function
Round 6 Second Stage: Moderate Difficulty
Test Information
θ

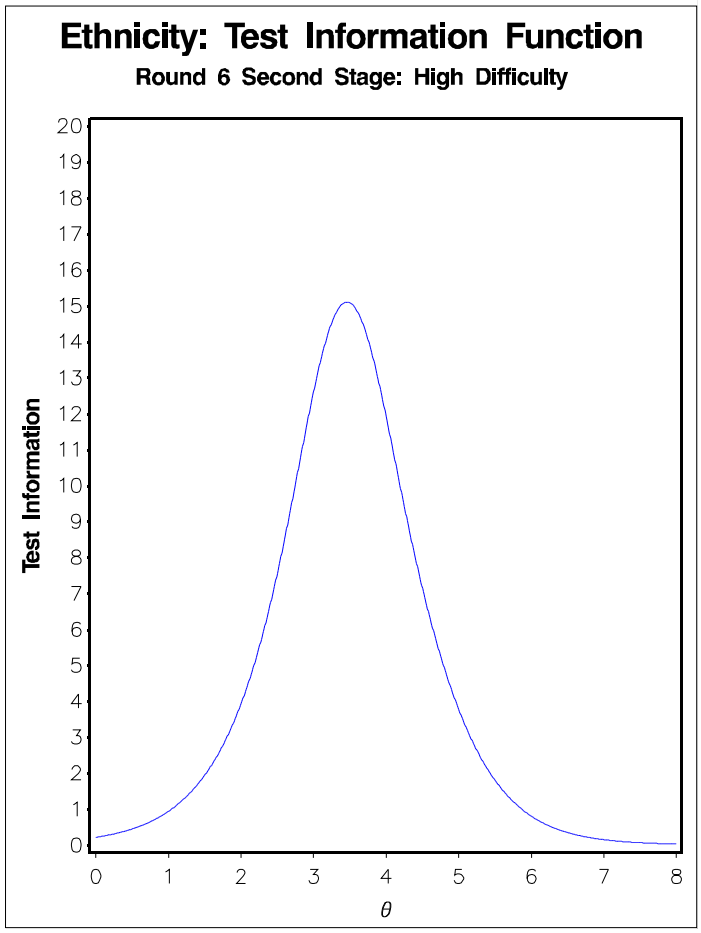
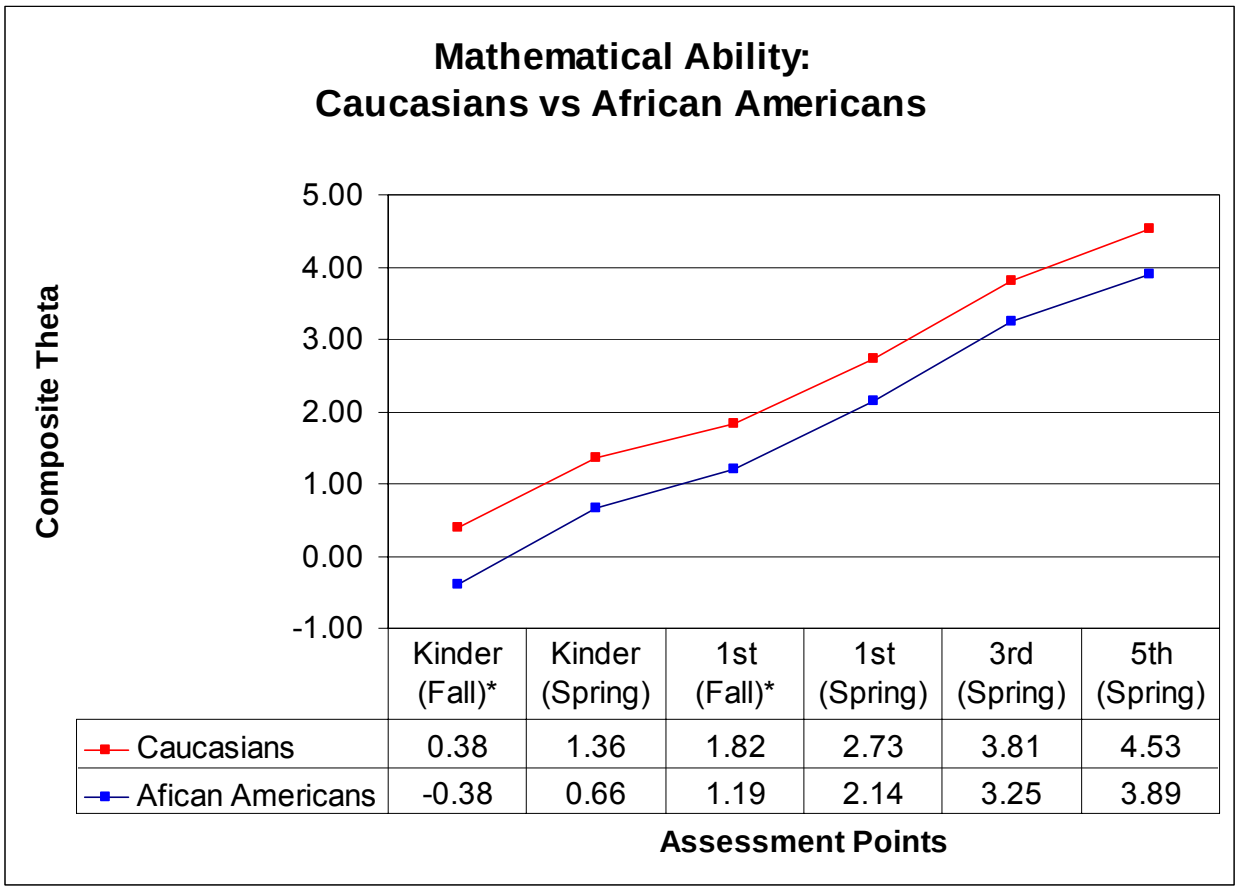

Ethnicity: Test Information Function
Round 6 Second Stage: High Difficulty
Test Information
θ

Figure 14

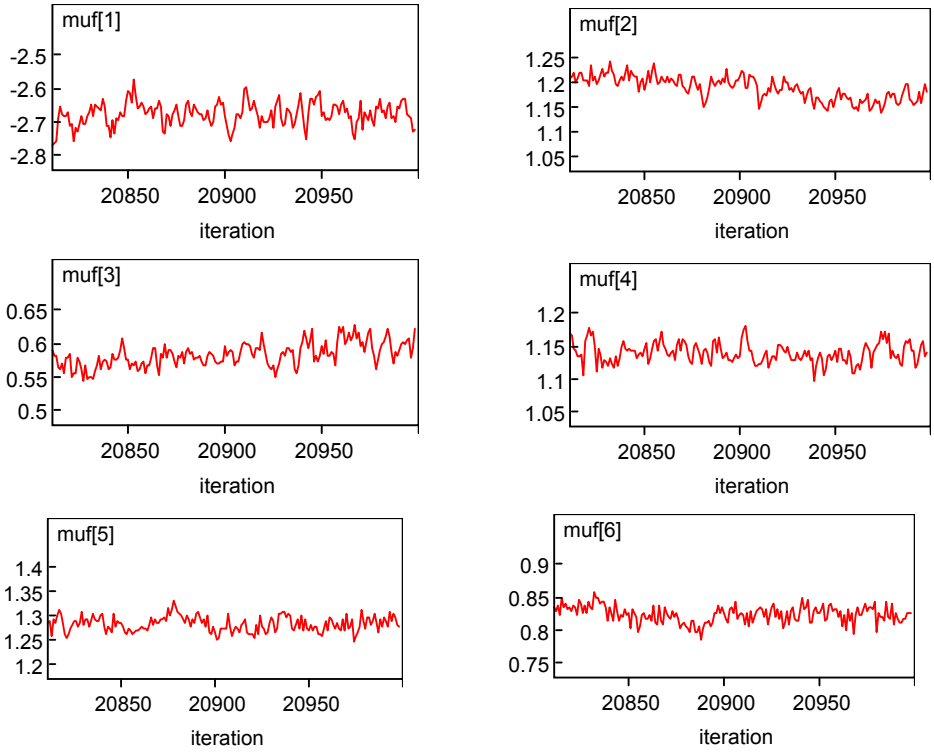


	Kinder (Fall)*	Kinder (Spring)	1st (Fall)*	1st (Spring)	3rd (Spring)	5th (Spring)
Caucasians	0.38	1.36	1.82	2.73	3.81	4.53
Afican Americans	-0.38	0.66	1.19	2.14	3.25	3.89

Note: * indicates statistically significant differences between the two groups.

Figure 15

Females



Males

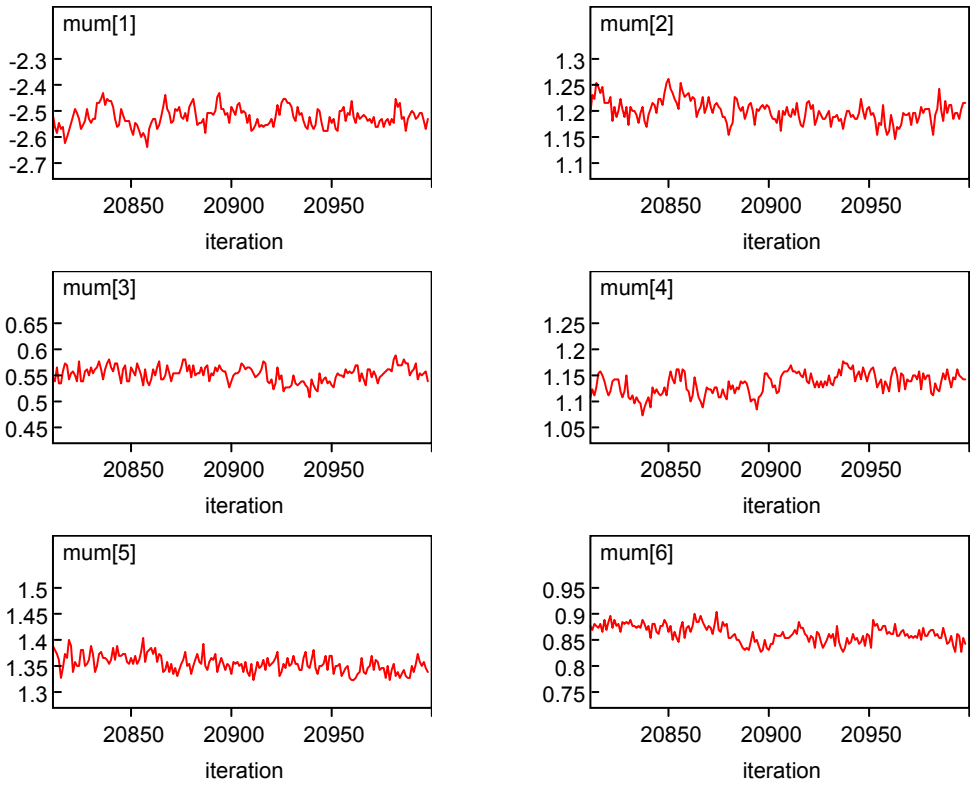
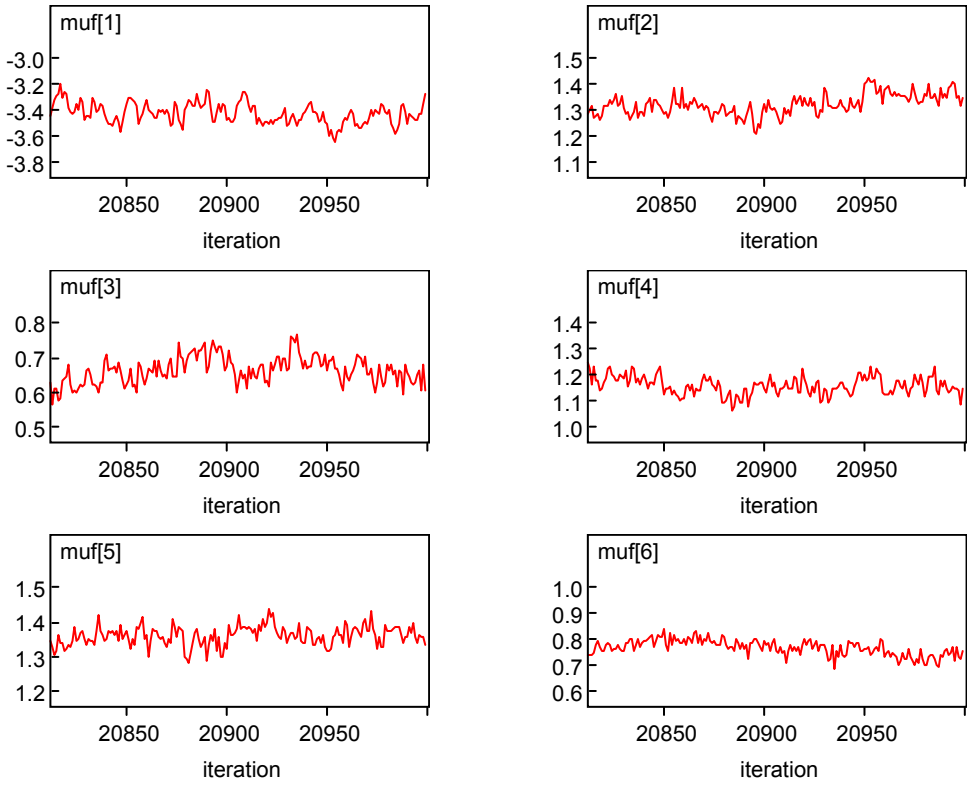


Figure 16

African Americans



Caucasians

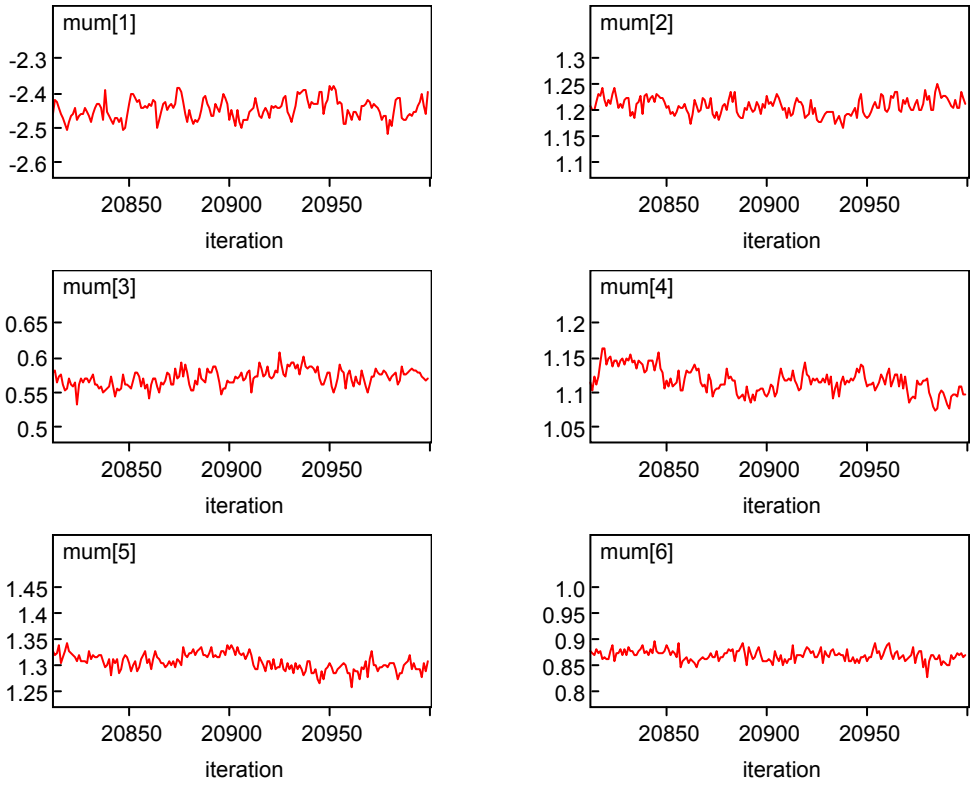
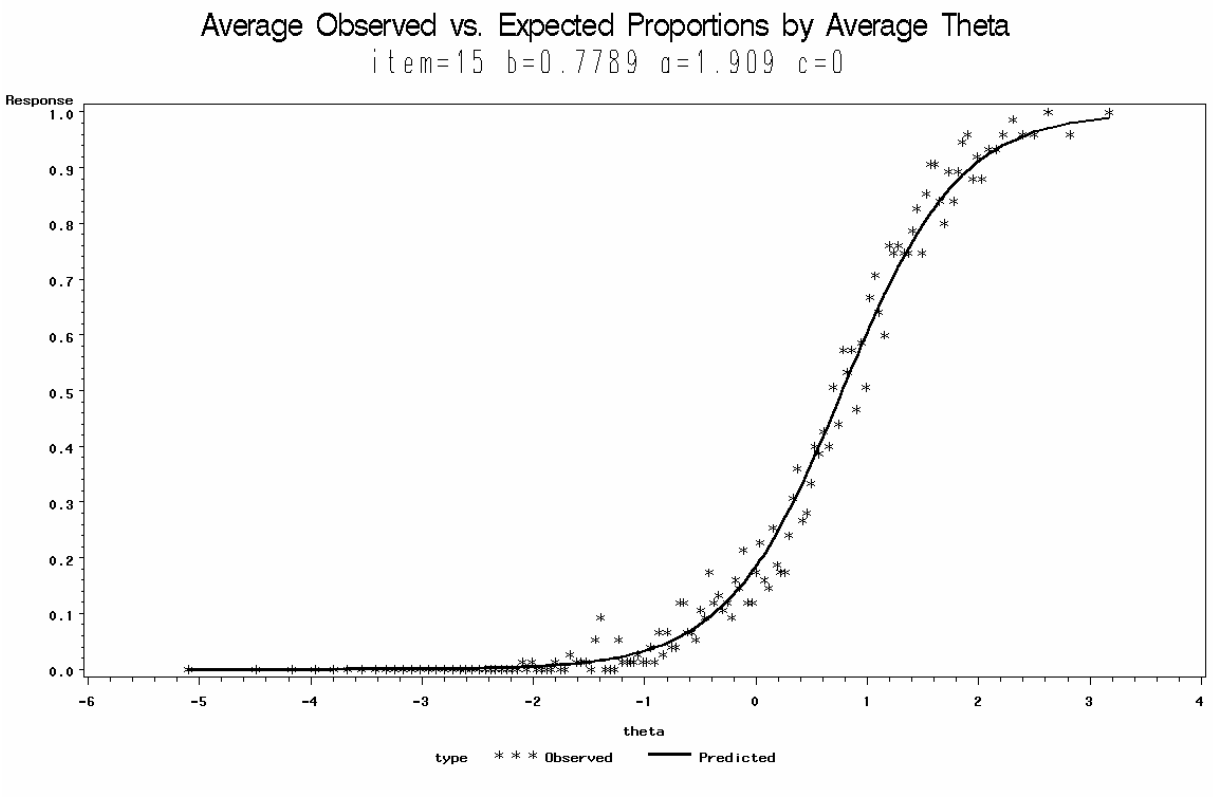


Figure 17


Average Observed vs. Expected Proportions by Average Theta
item=15 b=0.7789 a=1.909 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
theta
type * * * Observed Predicted

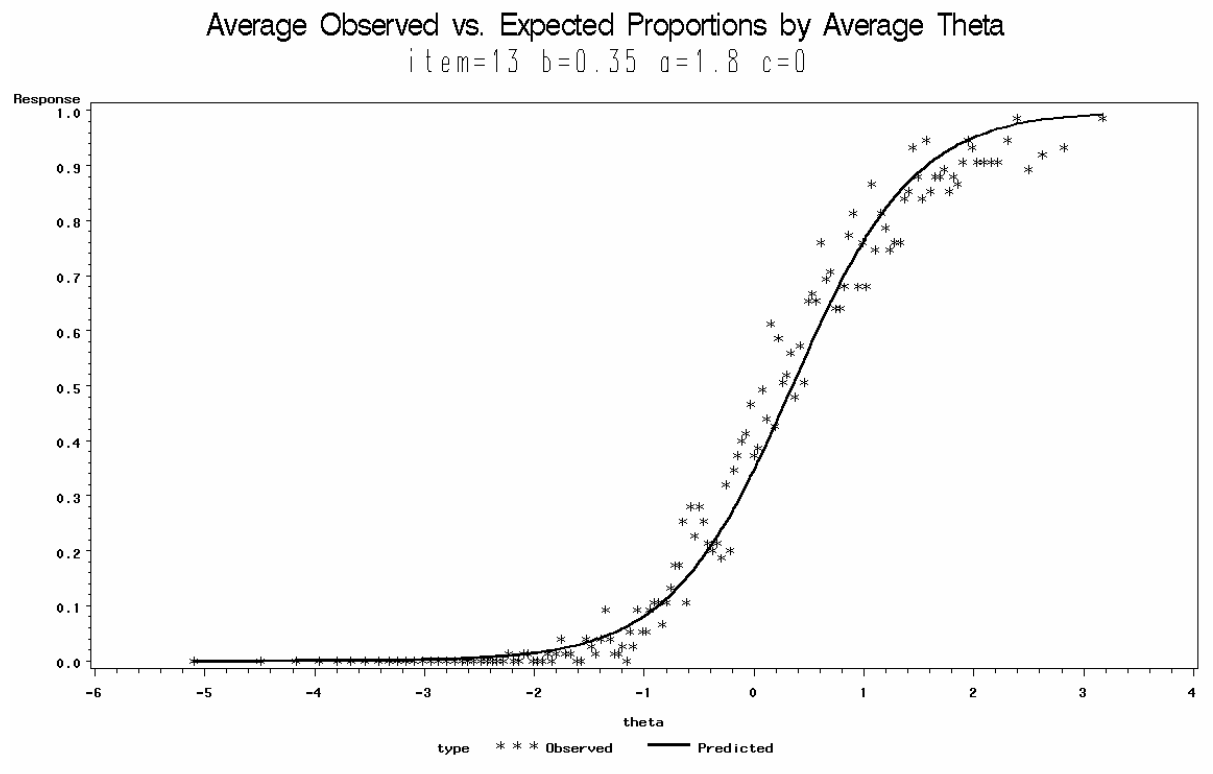
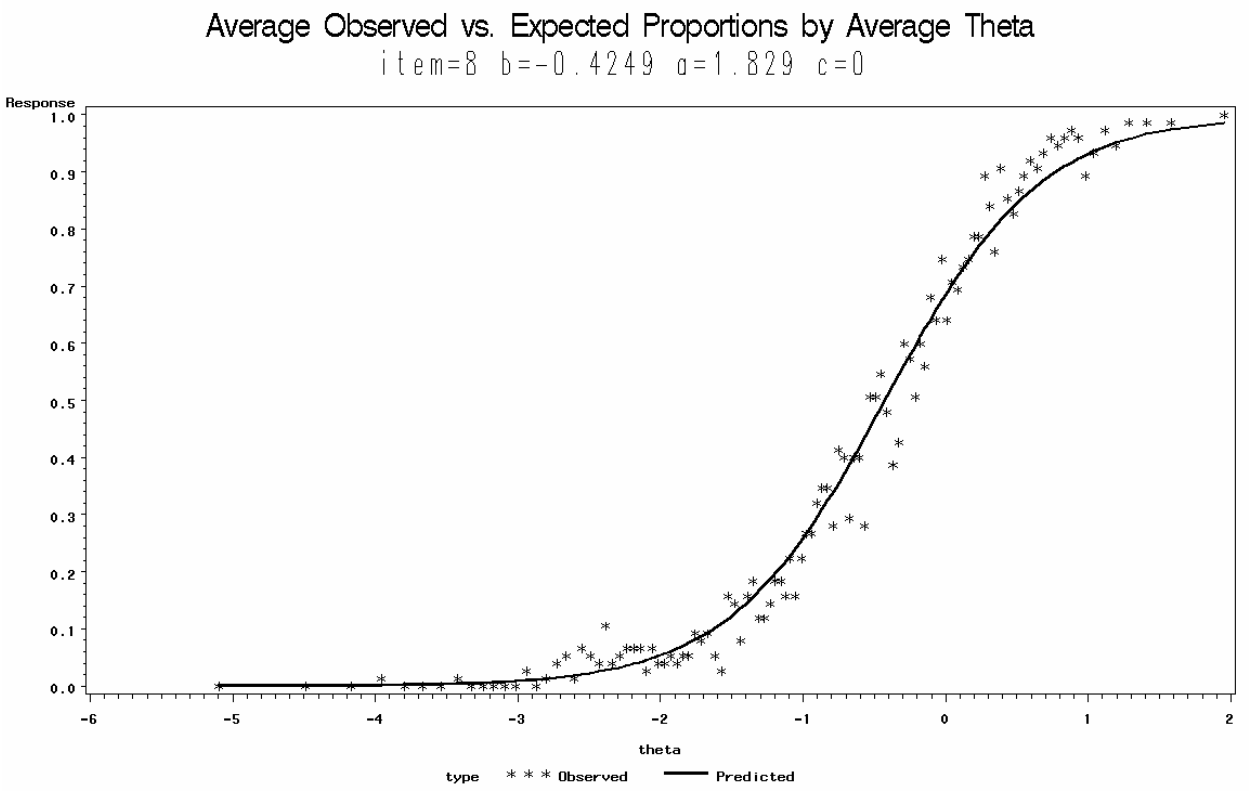

Average Observed vs. Expected Proportions by Average Theta
item=13 b=0.35 a=1.8 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
theta
type * * * Observed Predicted

Figure 18


Average Observed vs. Expected Proportions by Average Theta
item=8 b=-0.4249 a=1.829 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
theta
type * * * Observed Predicted

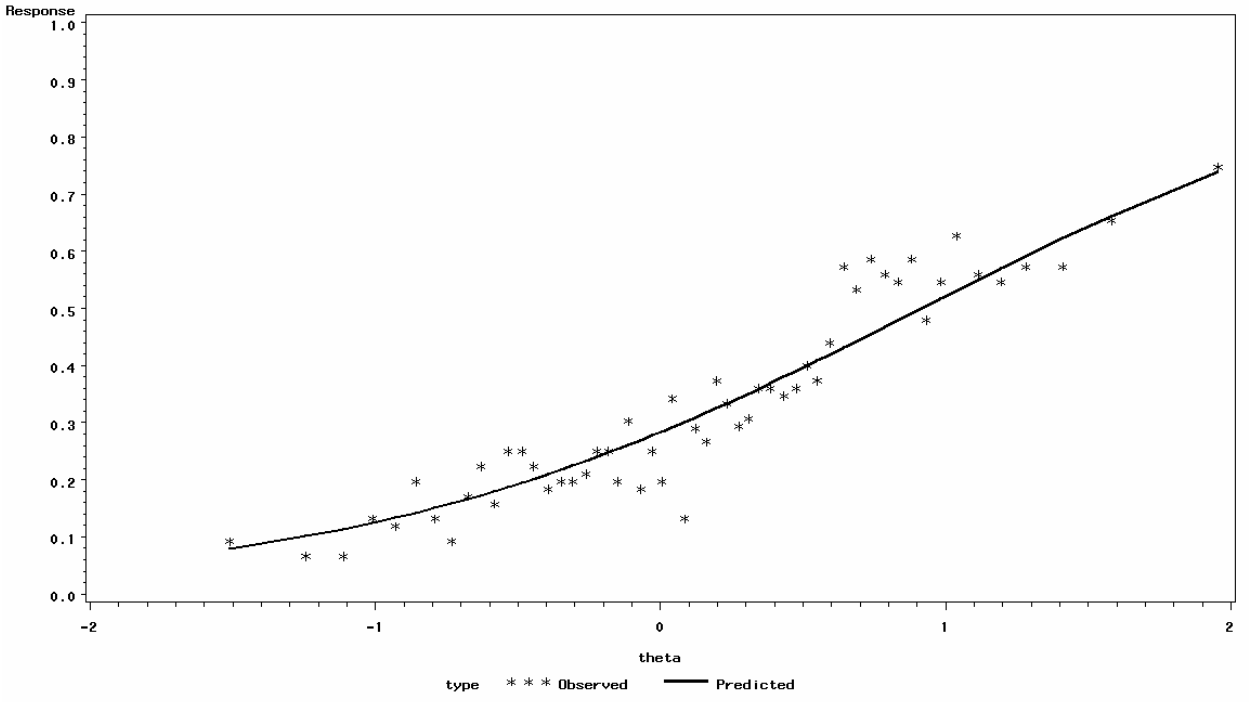
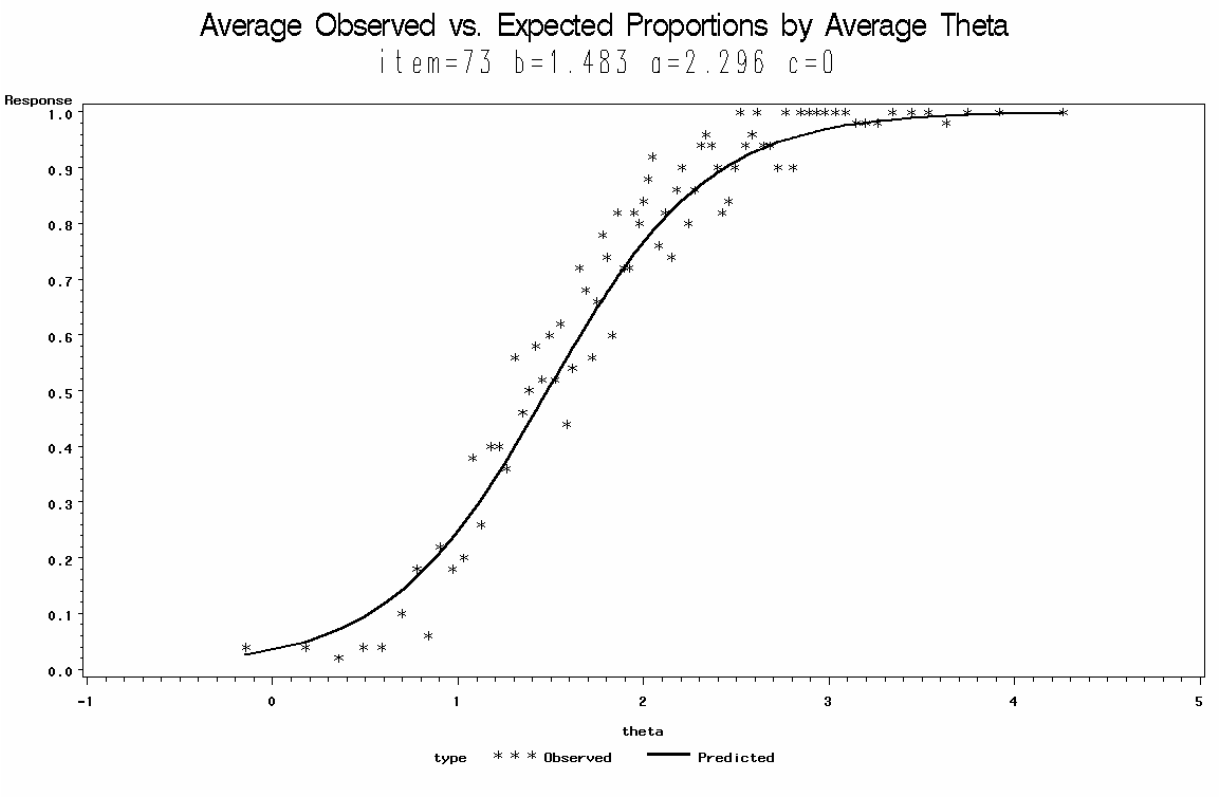

Average Observed vs. Expected Proportions by Average Theta
item=48 b=0.9175 a=1.013 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
theta
type * * * Observed Predicted

Figure 19


Average Observed vs. Expected Proportions by Average Theta
item=73 b=1.483 a=2.296 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-1
0
1
2
3
4
5
theta
type * * * Observed Predicted

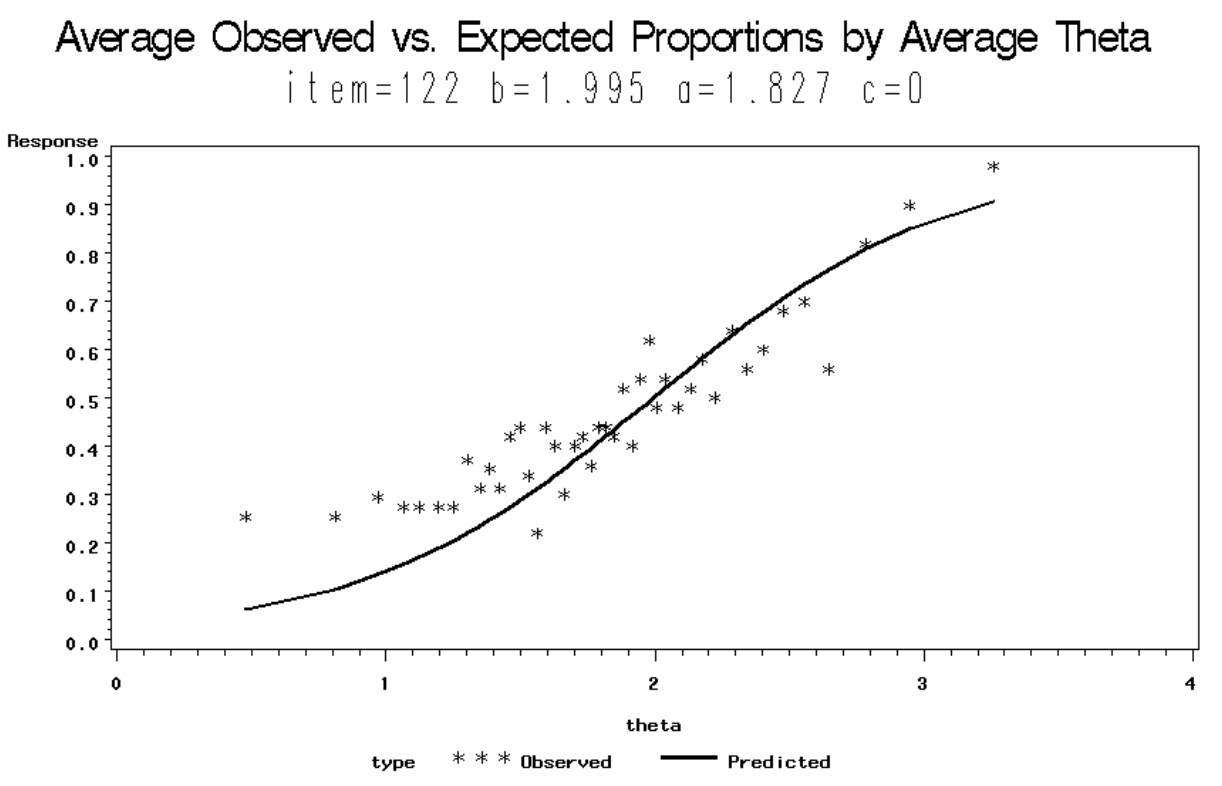
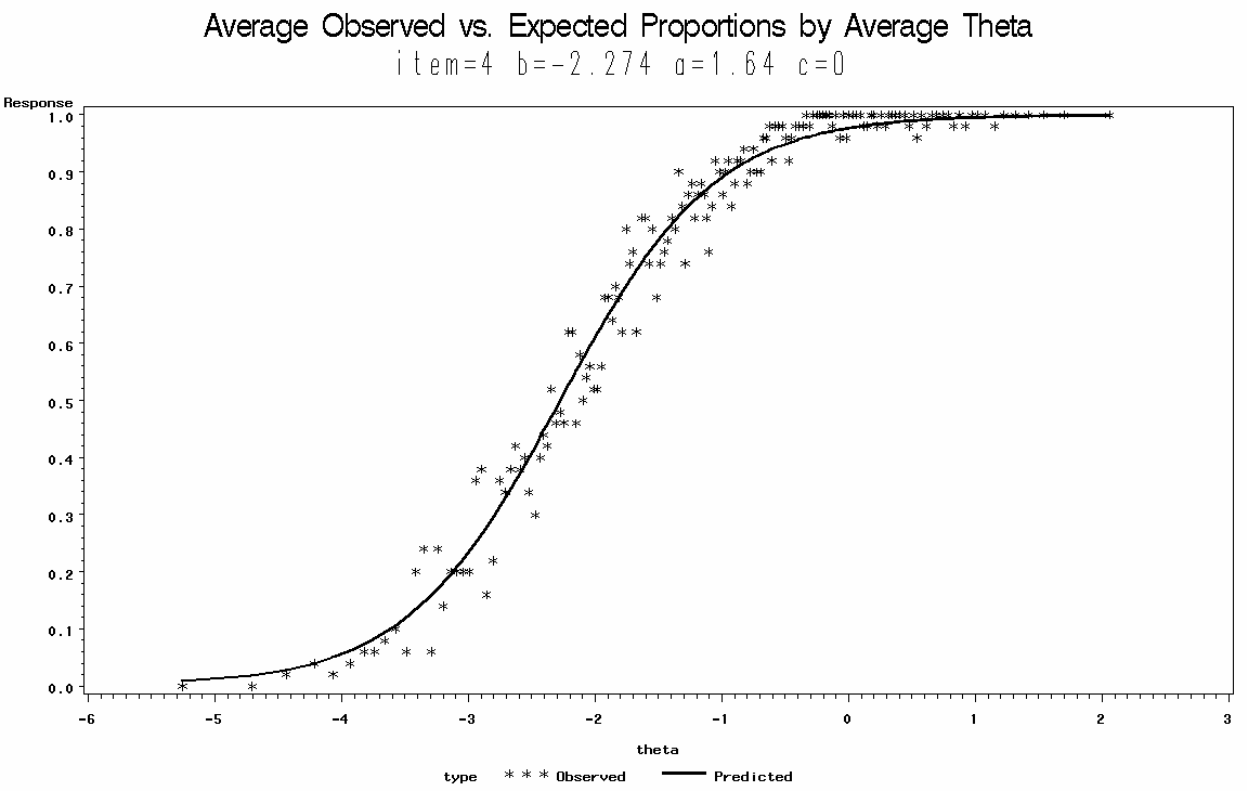

Average Observed vs. Expected Proportions by Average Theta
item=122 b=1.995 a=1.827 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
theta
type * * * Observed Predicted

Figure 20


Average Observed vs. Expected Proportions by Average Theta
item=4 b=-2.274 a=1.64 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
theta
type * * * Observed Predicted

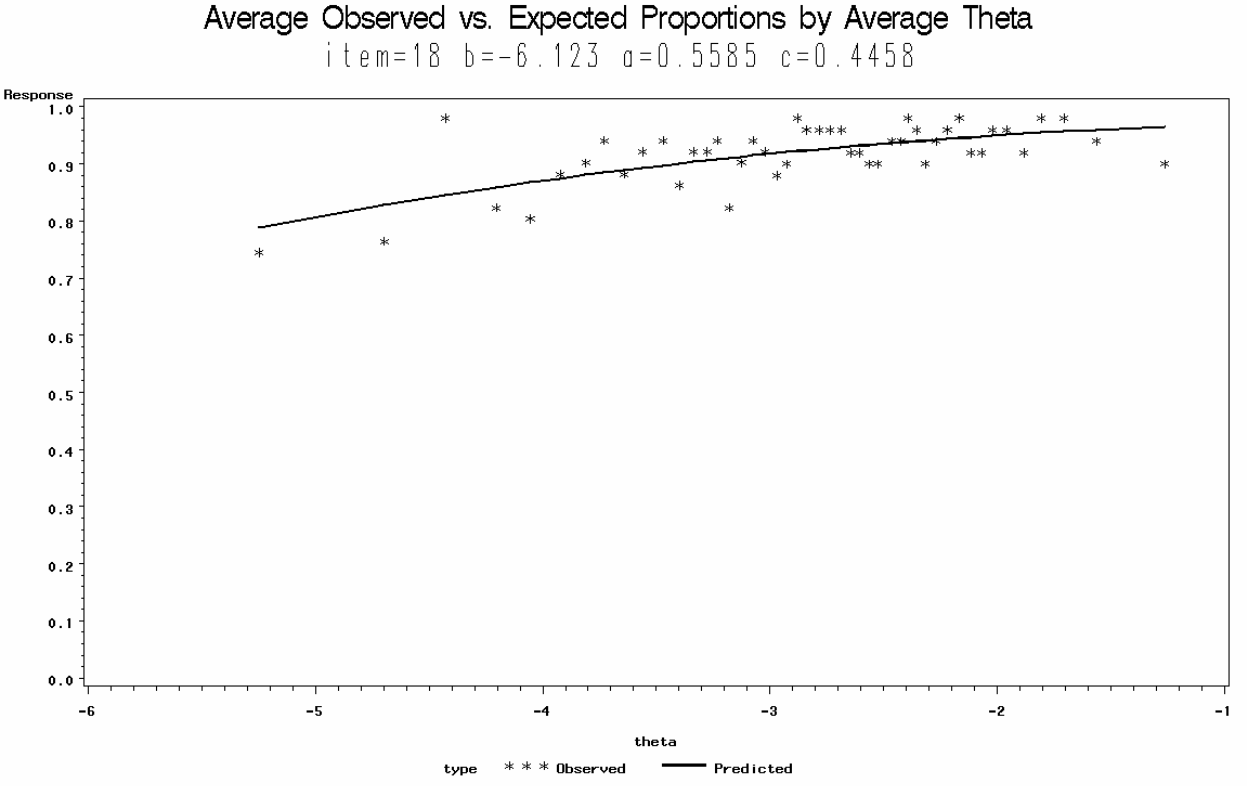
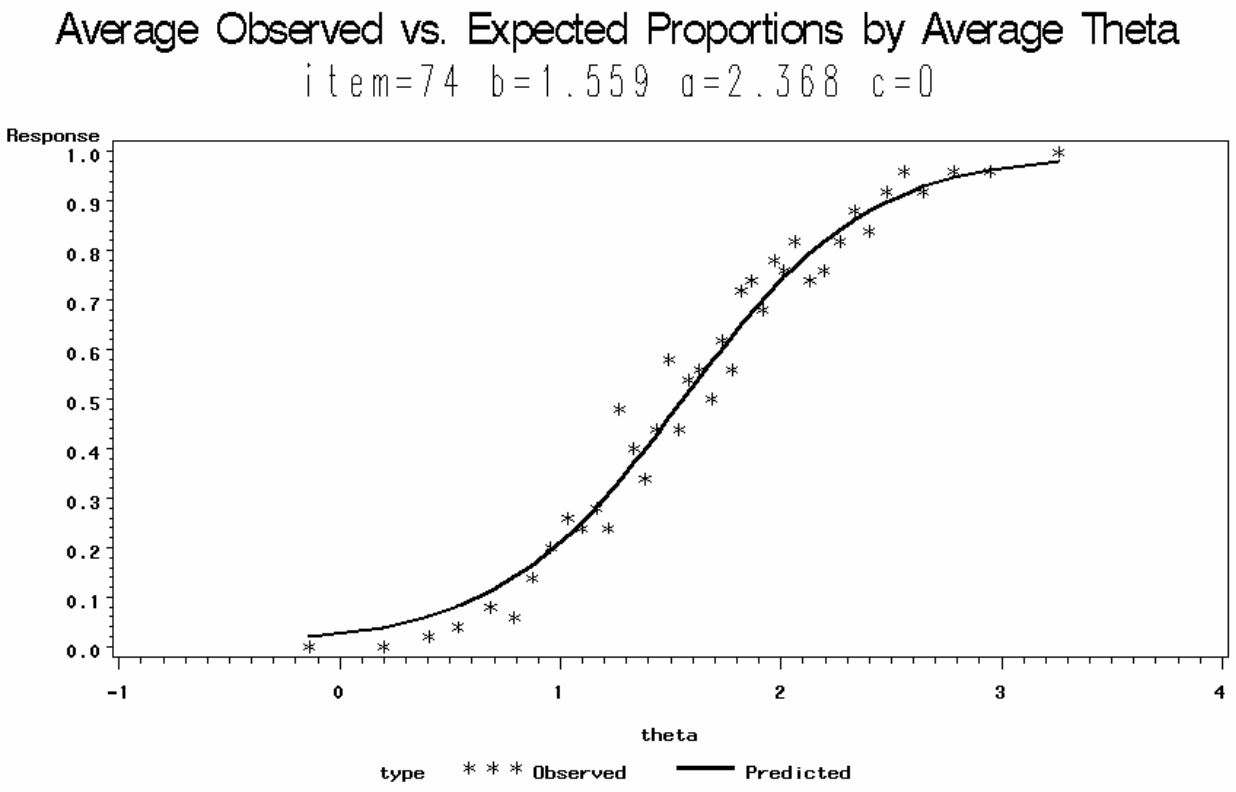

Average Observed vs. Expected Proportions by Average Theta
item=18 b=-6.123 a=0.5585 c=0.4458
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
theta
type * * * Observed Predicted

Figure 21


Average Observed vs. Expected Proportions by Average Theta
item=74 b=1.559 a=2.368 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-1
0
1
2
3
4
theta
type * * * Observed Predicted

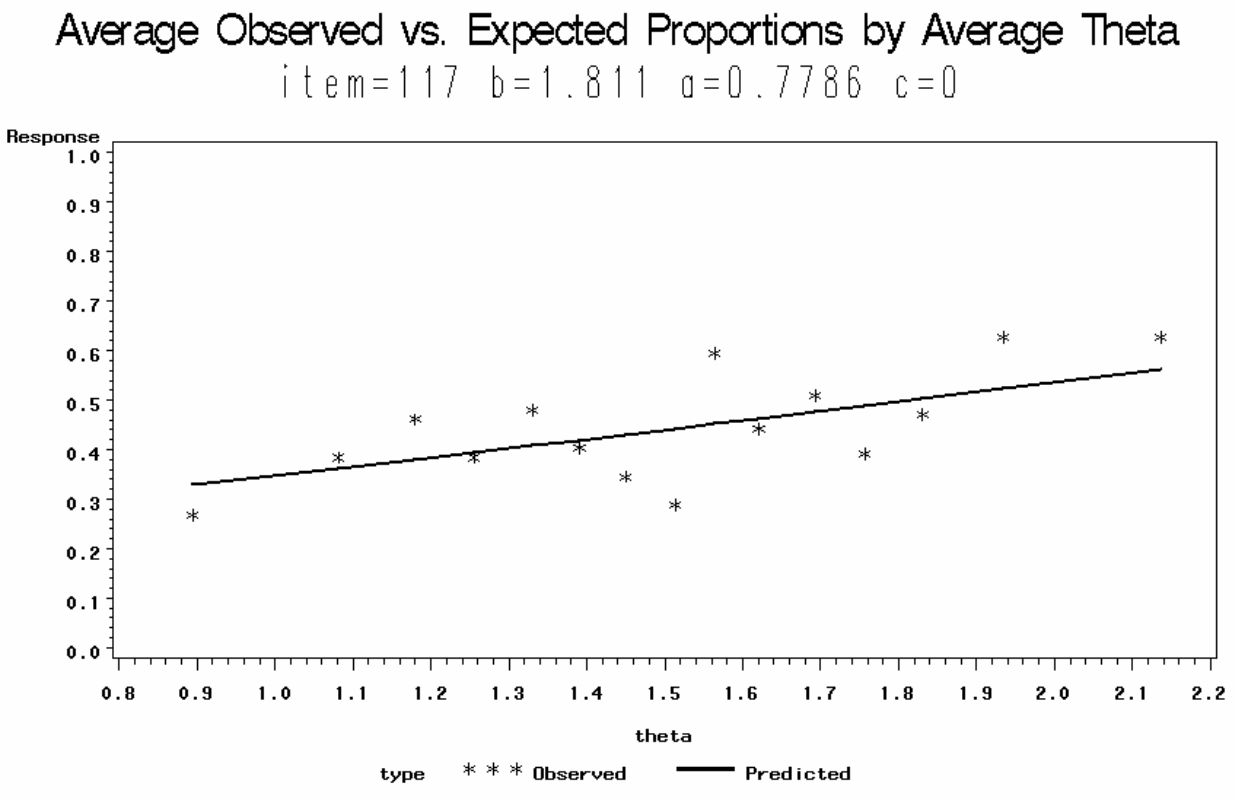
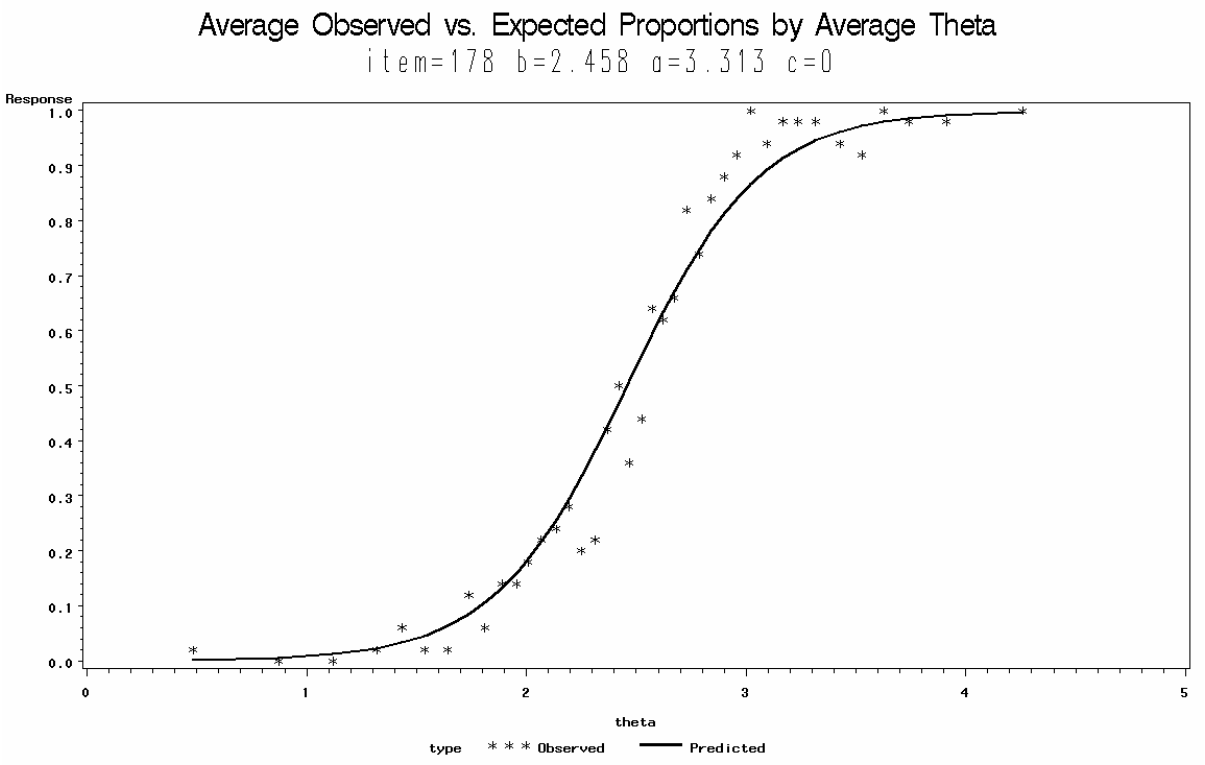

Average Observed vs. Expected Proportions by Average Theta
item=117 b=1.811 a=0.7786 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
theta
type * * * Observed Predicted

Figure 22


Average Observed vs. Expected Proportions by Average Theta
item=178 b=2.458 a=3.313 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
theta
type * * * Observed Predicted

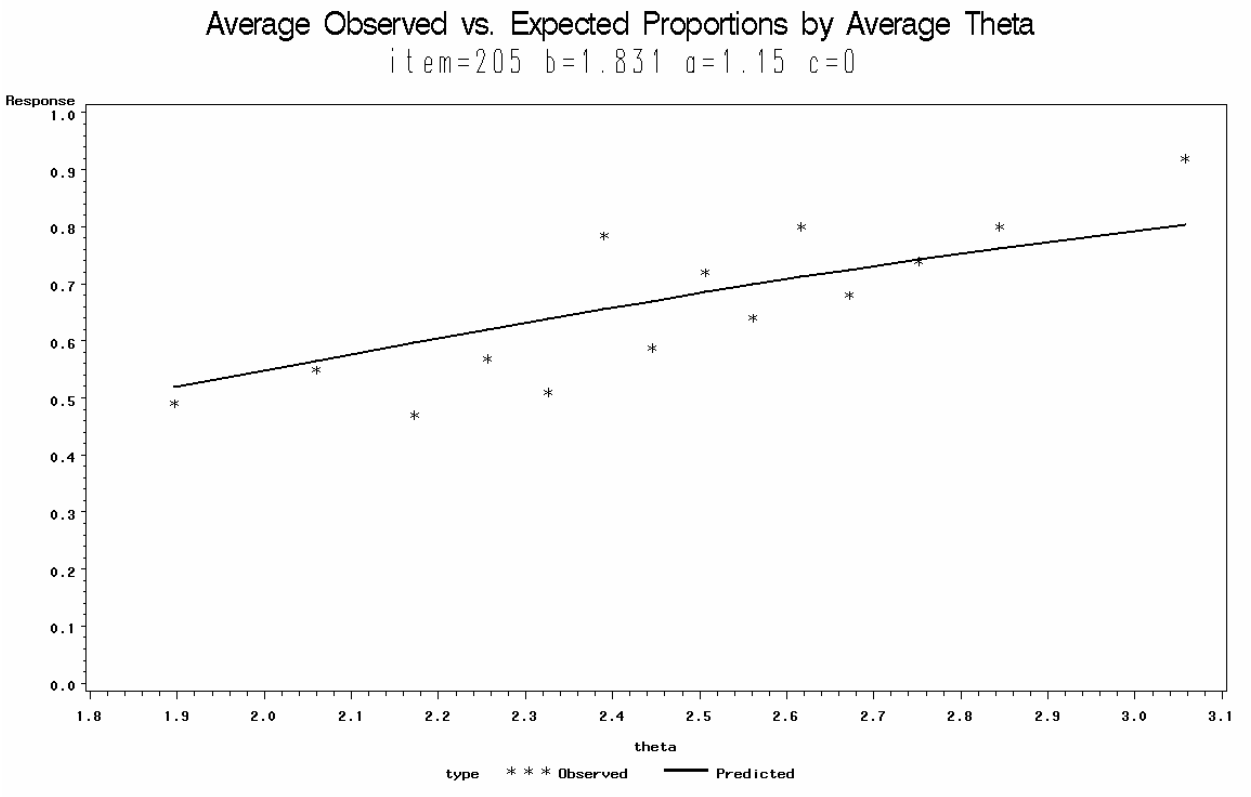
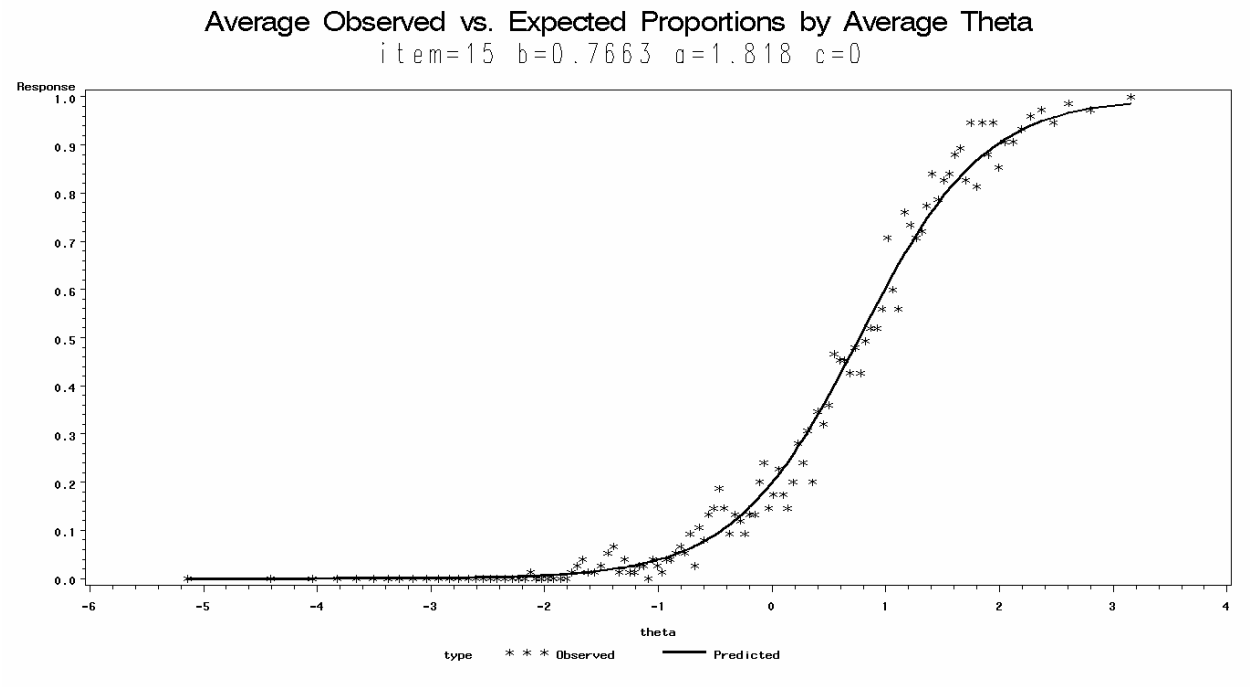

Average Observed vs. Expected Proportions by Average Theta
item=205 b=1.831 a=1.15 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
theta
type * * * Observed Predicted

Figure 23


Average Observed vs. Expected Proportions by Average Theta
item=15 b=0.7663 a=1.818 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
theta
type * * * Observed Predicted

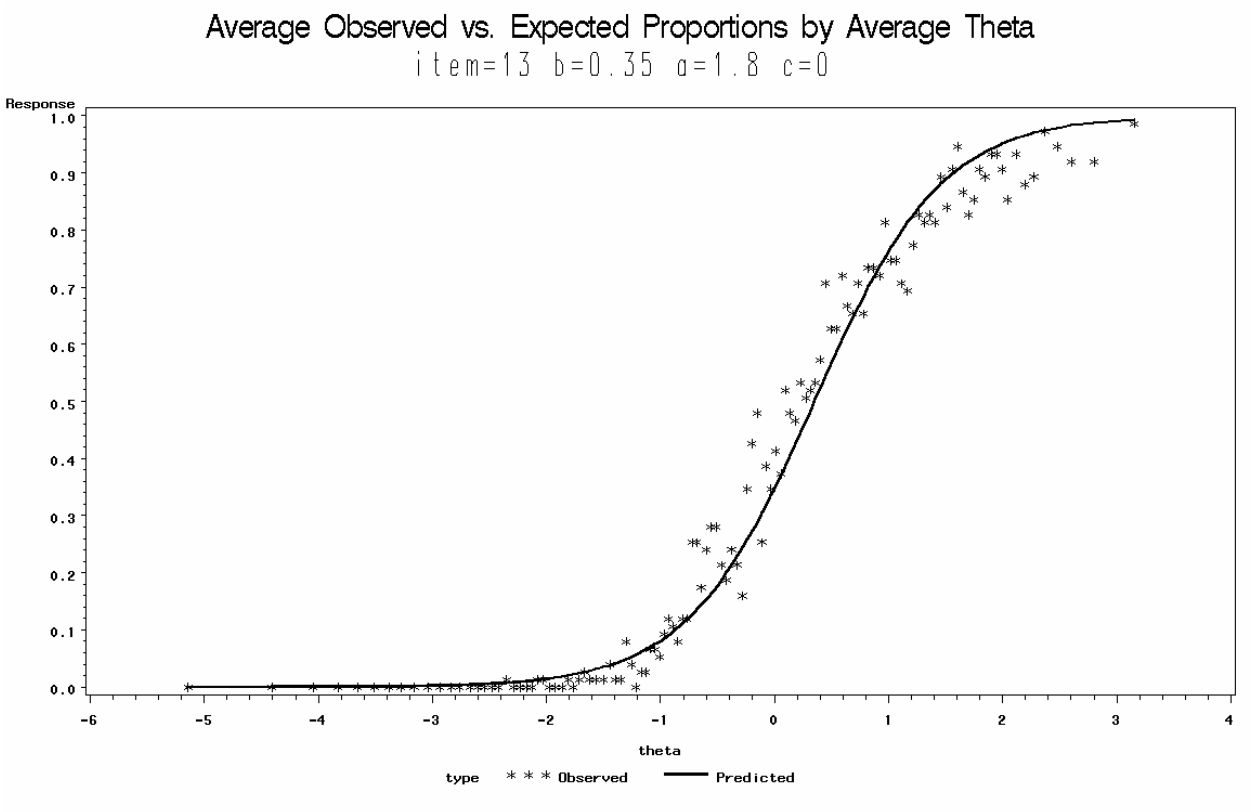
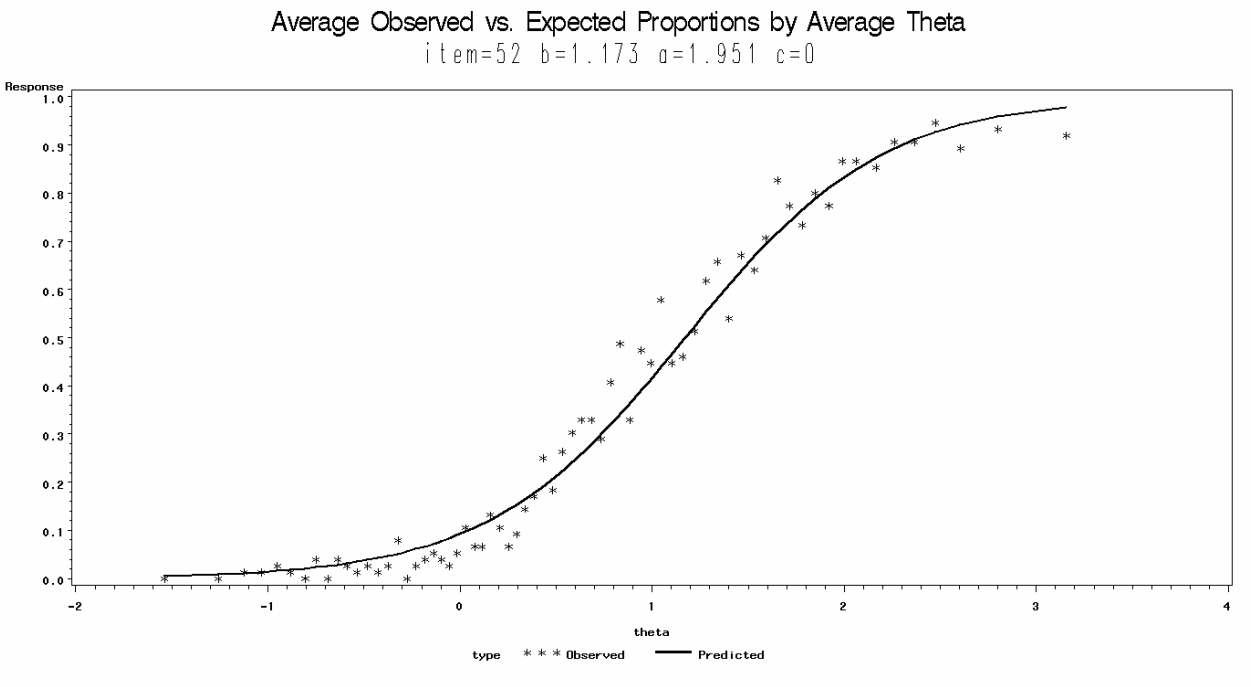

Average Observed vs. Expected Proportions by Average Theta
item=13 b=0.35 a=1.8 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
theta
type * * * Observed Predicted

Figure 24


Average Observed vs. Expected Proportions by Average Theta
item=52 b=1.173 a=1.951 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
3
4
theta
type * * * Observed Predicted

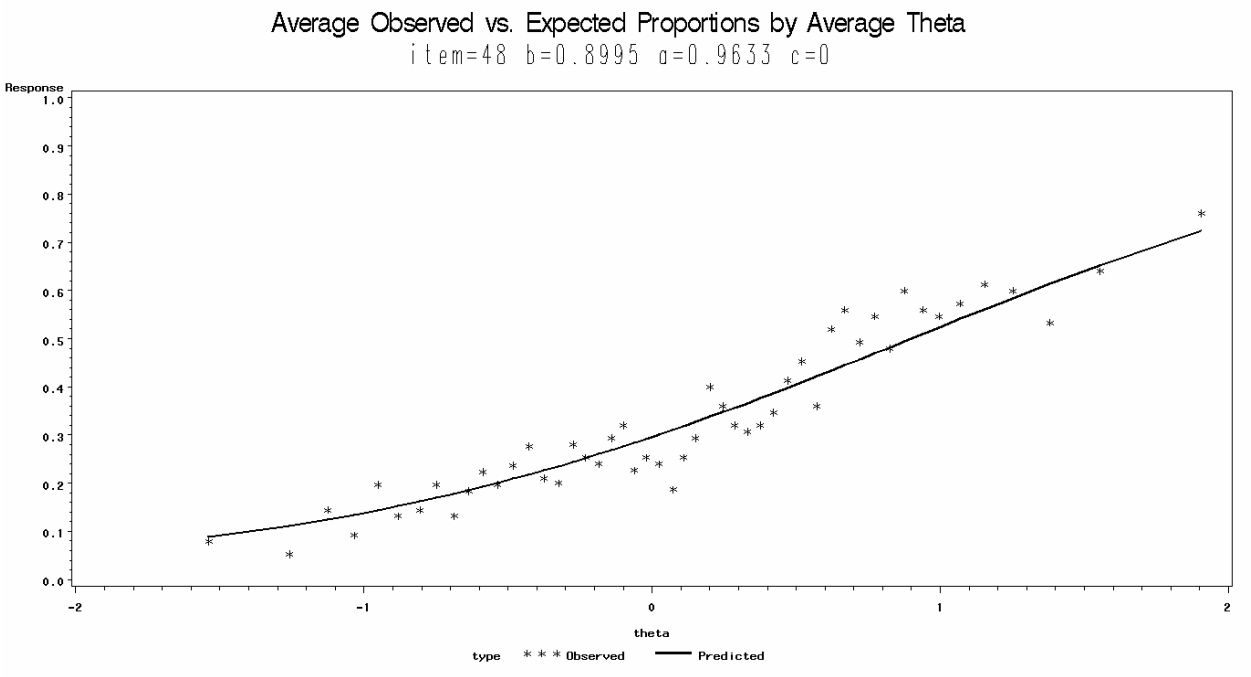
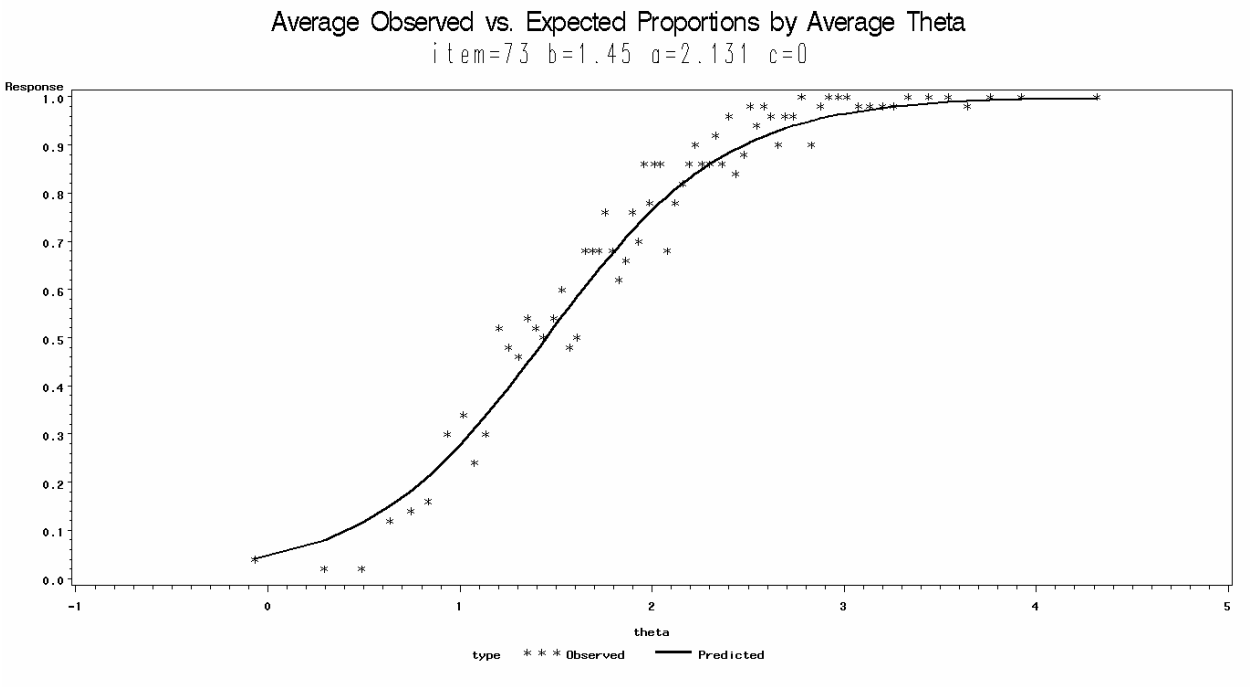

Average Observed vs. Expected Proportions by Average Theta
item=48 b=0.8995 a=0.9633 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
theta
type * * * Observed Predicted

Figure 25


Average Observed vs. Expected Proportions by Average Theta
item=73 b=1.45 a=2.131 c=0
Response
theta
type * * * Observed Predicted

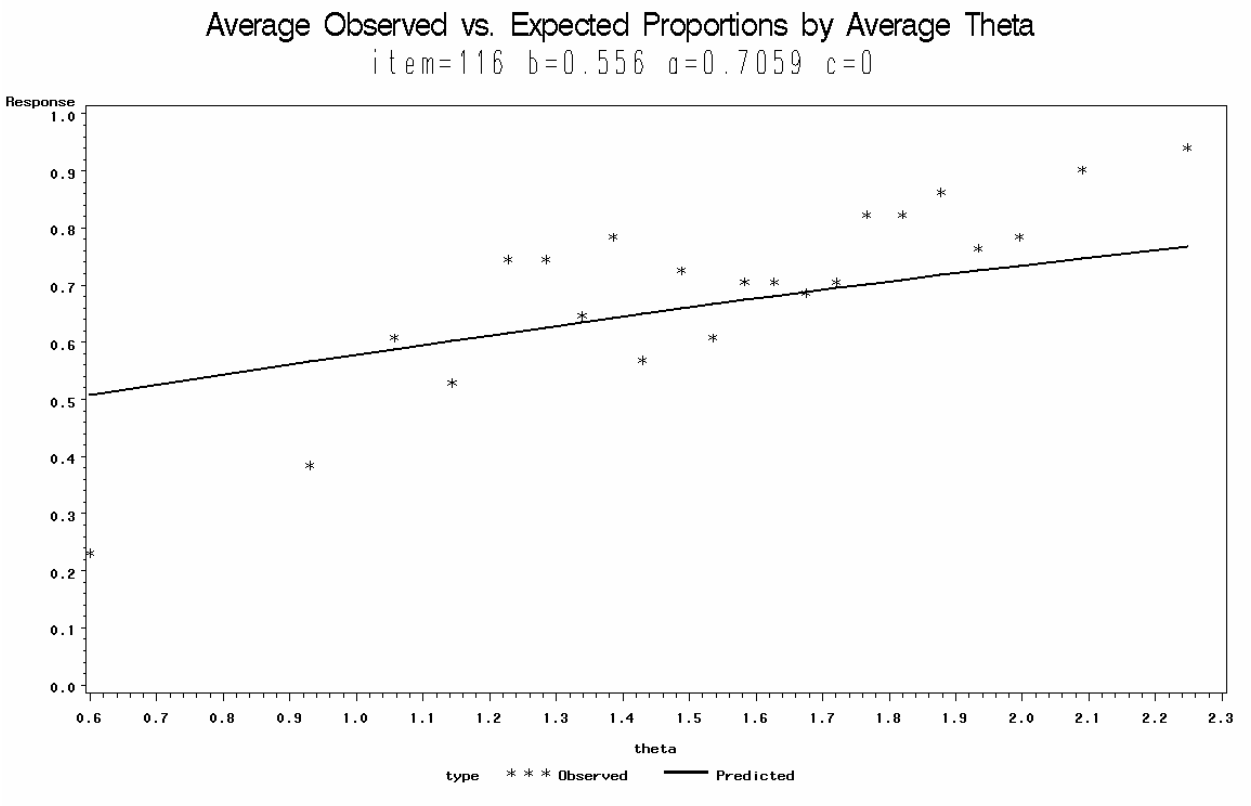
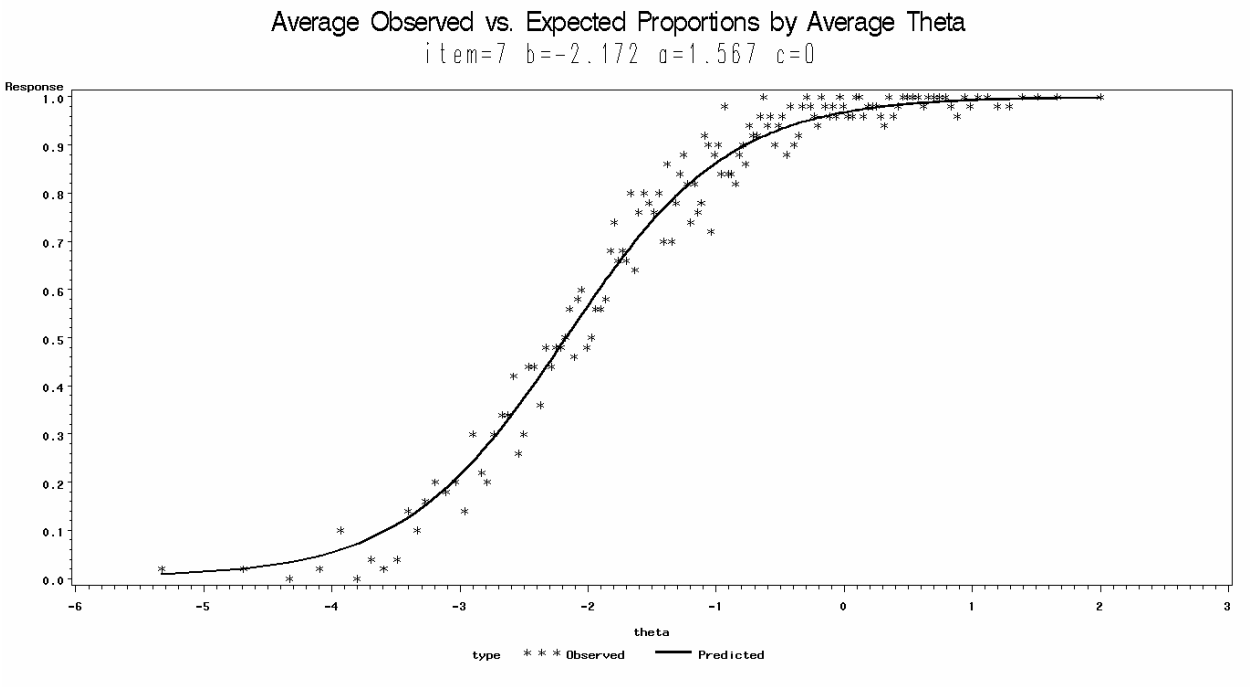

Average Observed vs. Expected Proportions by Average Theta
item=116 b=0.556 a=0.7059 c=0
Response
theta
type * * * Observed Predicted

Figure 26


Average Observed vs. Expected Proportions by Average Theta
item=7 b=-2.172 a=1.567 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
0
1
2
3
theta
type * * * Observed Predicted

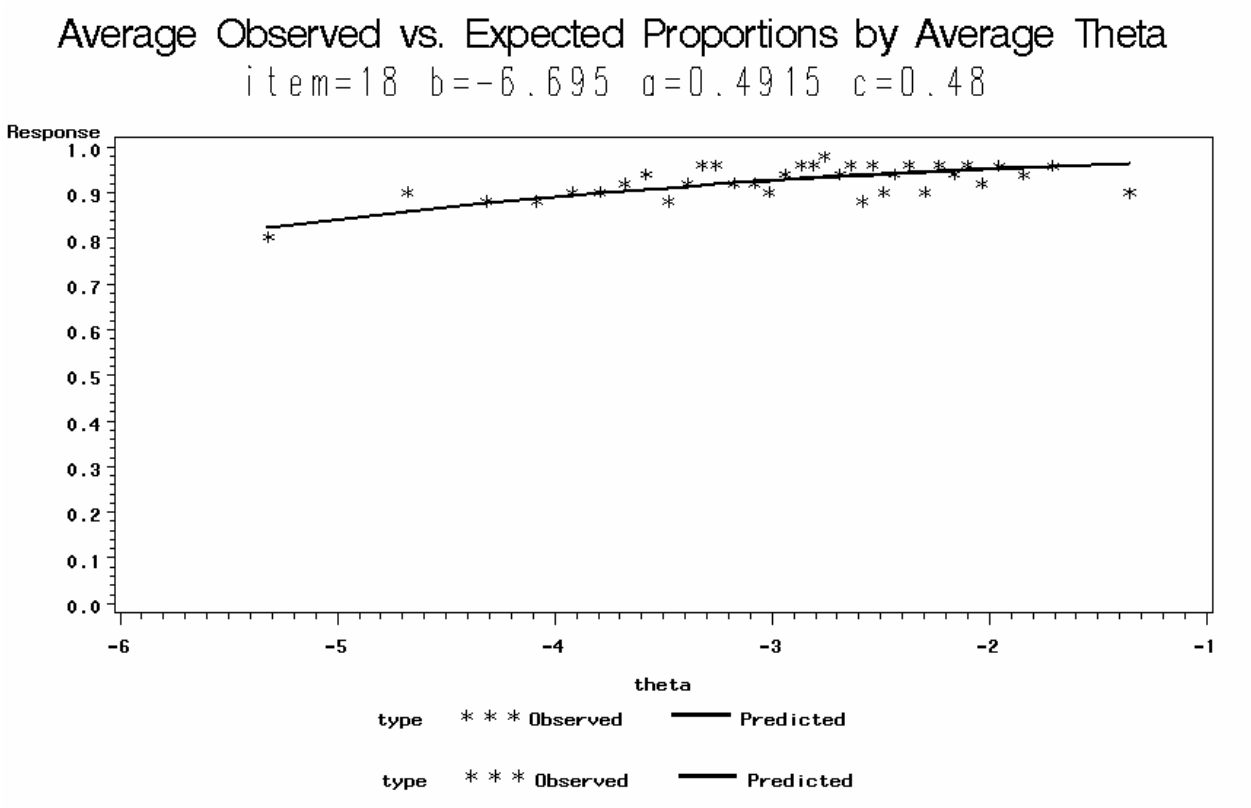
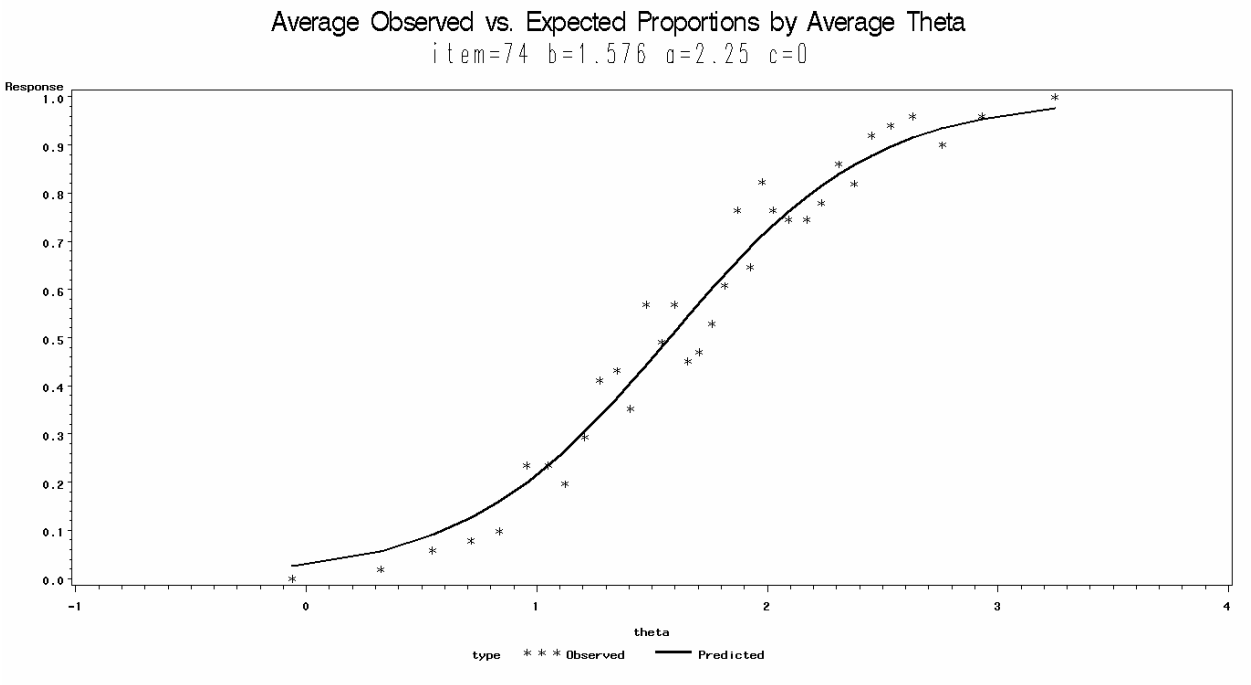

Average Observed vs. Expected Proportions by Average Theta
item=18 b=-6.695 a=0.4915 c=0.48
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-6
-5
-4
-3
-2
-1
theta
type * * * Observed Predicted
type * * * Observed Predicted

Figure 27


Average Observed vs. Expected Proportions by Average Theta
item=74 b=1.576 a=2.25 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-1
0
1
2
3
4
theta
type * * * Observed Predicted

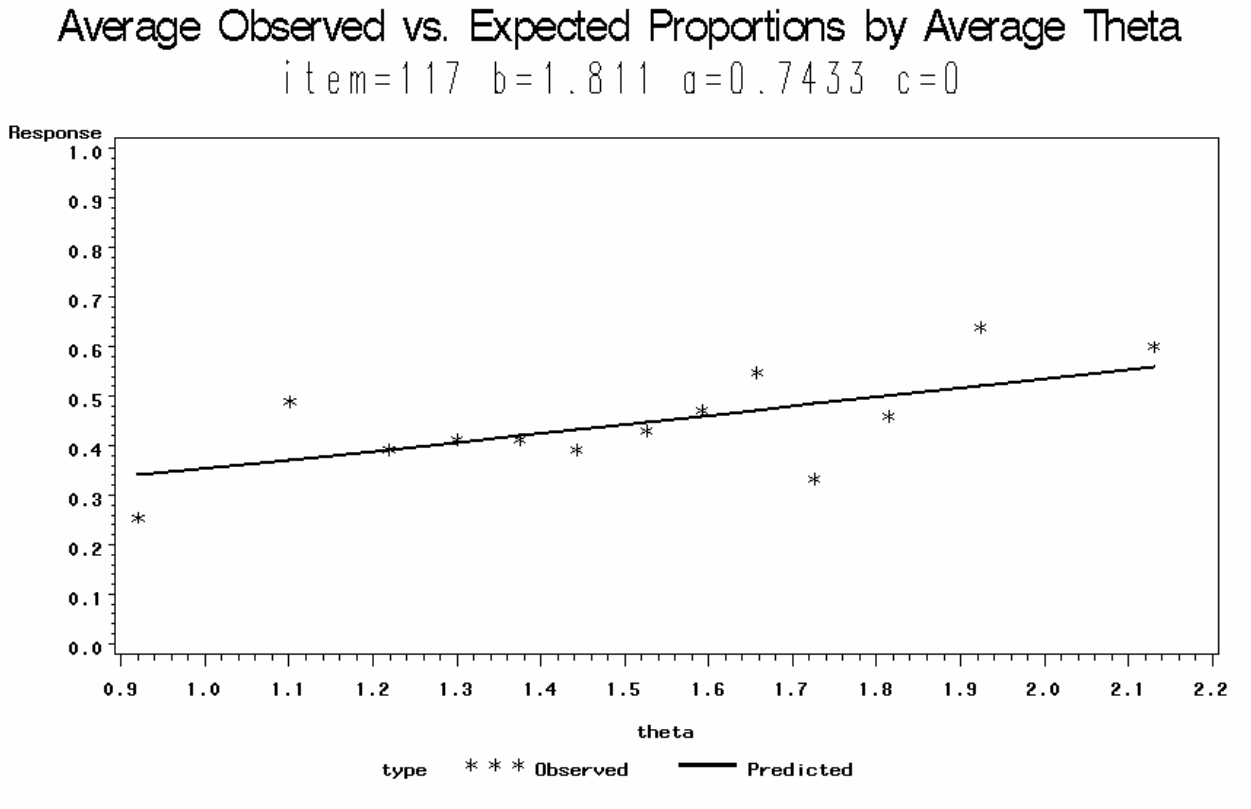
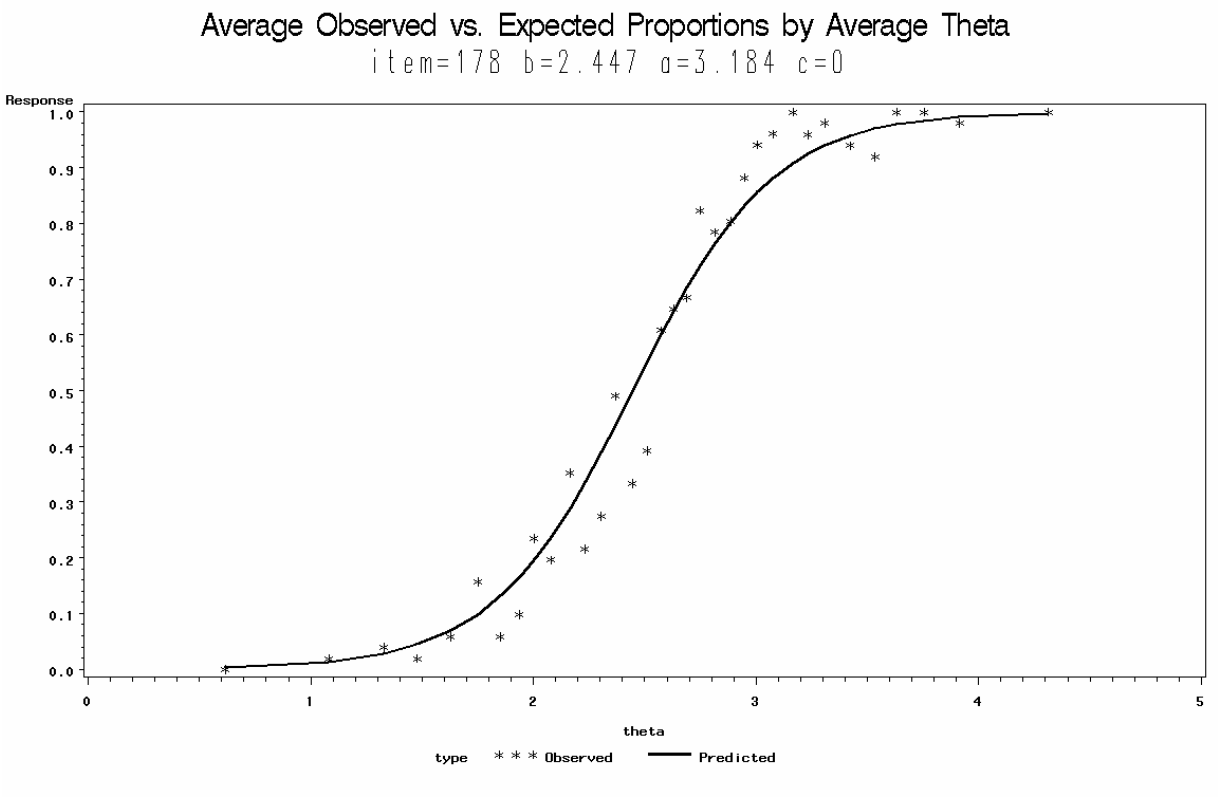
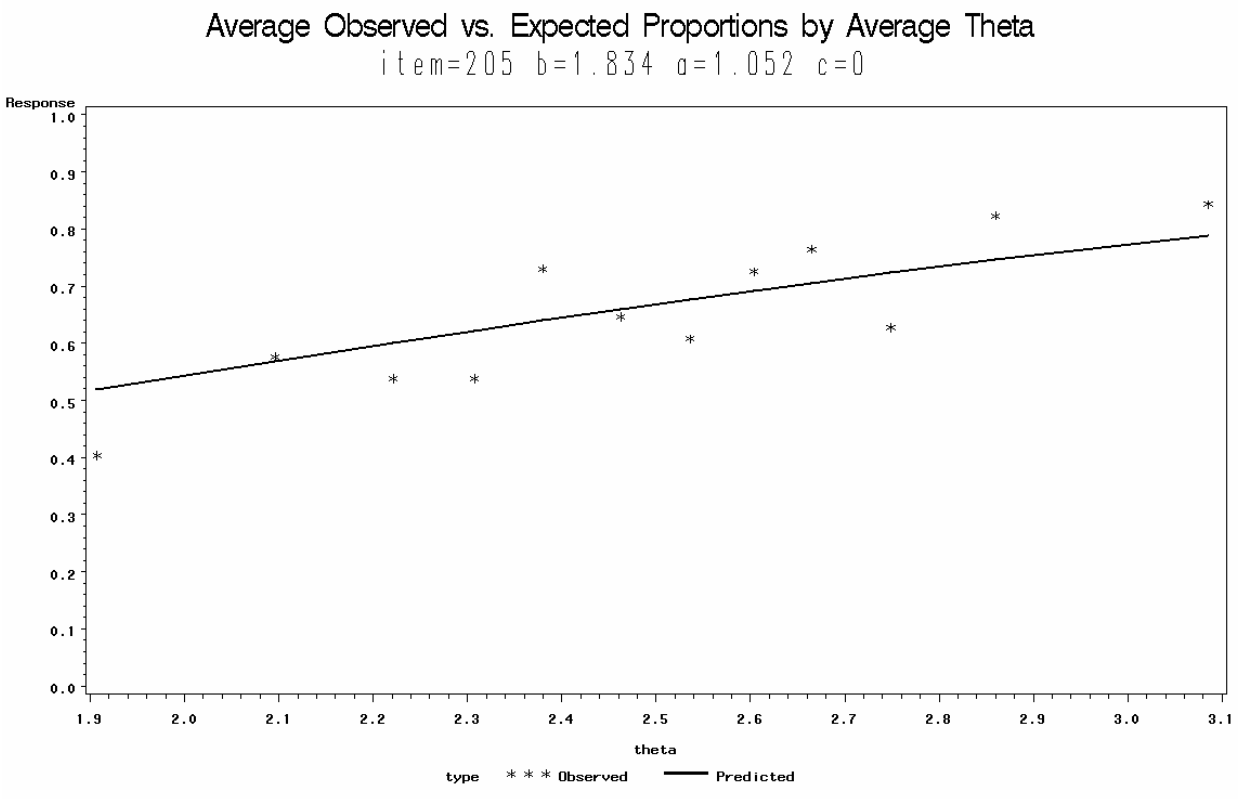

Average Observed vs. Expected Proportions by Average Theta
item=117 b=1.811 a=0.7433 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
theta
type * * * Observed Predicted

Figure 28


Average Observed vs. Expected Proportions by Average Theta
item=178 b=2.447 a=3.184 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
theta
type * * * Observed Predicted


Average Observed vs. Expected Proportions by Average Theta
item=205 b=1.834 a=1.052 c=0
Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
theta
type * * * Observed Predicted

References

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

Aunola, K., Leskinene, E., Lerkkanen, M.K., & Nurmi, J.E. (2004). Developmental dynamics of math performance from preschool to grade 2. Journal of Educational Psychology, 96(4), 699-713.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), Problems in measuring change (pp.3-20). Madison, WI: University of Wisconsin Press.

Chandler, K., West, J., & Hausken, E. (1995). Approaching Kindergarten: A look at preschoolers in the United States. (NCES Report No. 95280). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Dimitrov, M. D., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. Work, 20, 159-165.

Donoghue, J.R., & Isham, S.P. (1998). A comparison of procedures to detect item parameter drift. Applied Psychological Measurement, 22(1), 33- 51.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56 (3), 495-515.

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

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. L. Linn (ed.), Educational measurement (pp. 105-146). New York: Macmillan.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. Applied Psychological Measurement, 27(1), 3-26.

Gluck, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. Methods of Psychological Research Online, 2. Retrieved August 9, 2007, from http://www.mpr-online.de

Hyde, J.S., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107(2), 139-155.

Jodoin, M.G., Keller, L.A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. The Journal of Experimental Education, 71(3), 229-250.

Kim, S.H., & Cohen, A.S. (1998). A comparison of linking and concurrent calibration under item response theory. Applied psychological measurement, 22(2), 131-143.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 47(1), 121-150.

Lord, F. M. (1956). The measurement of growth. Educational and Psychological Measurement, 16(4), 421-437.

Lord, F.M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement, 18(3), 437-451.

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

Mellenbergh, G.J. (1999). A note on simple gain score precision. Applied Psychological Measurement, 23(1), 87-89.

Overall, J.E., & Woodward, J.A. (1975). Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 82(1), 85-86.

Rock, D.A., & Pollack, J.M. (2002). Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade. (Working Paper Series No. 2002-05). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Roberts, J.S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. Lissitz (Ed.), Longitudinal and value added modeling of student performance. Maple, MN: JAM Press.

Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of difference score in the measurement of change. Journal of Educational Measurement, 20(4), 335-343.

Rogosa, D.R., & Willett, J.B. (1985). Understanding correlates of change by modeling individual differences in growth. Psychometrika, 50(2), 203-228.

Sinharay, S. (2003). Assessing convergence of the Markov Chain Monte Carlo Algorithms: A Review. (Report No. RR-03-07). Princeton, NJ: Educational Testing Service.

Speigelhalter, D., Thomas, A., Best, N., & Dunn, D. (2007). WinBUGS user manual version 4.1.3. Cambridge: MRC Biostatistics Unit.

Speigelhalter, D., Best, N., Carlin, B., & Van der Linden, A. (2002). A Baeyesian Measures of Model of Complexity and Fit (with Discussion). Journal of the Royal Statistical Society, 64(4), 583-616.

Tate, W.F. (1997). Race ethnicity, SES, gender and language proficiency trends in Mathematics achievement: An update. Journal for Research in Mathematics Education, 28(6), 652-79.

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

Thorndike, E. (1924). The influence of change imperfections of measures upon the relation of initial score to gain or loss. Journal of Experimental Psychology, 7, 225-232.

Wang, W.C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. Educational and Psychological Measurement, 64(6), 937-955.

Wang, W.C., Chen, P.H., & Cheng, Y.Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. Psychological Methods, 9(1), 116-136.

Wang, W., Wilson, M., & Adams, R. (1998). Measuring individual differences in change with multidimensional Rasch models. Journal of Outcome Measurement, 2(3), 240-265.

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. Educational and Psychological Measurement, 37(3), 679-689.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? Applied Psychological measurement, 20(1), 59-69.

Willett, J. B. (1989). Questions and answers in the measurement of change. In E.Z. Rothkopf (Ed.), Review of research in education (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.

Willett, J.B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. Educational and Psychological Measurement, 49(3), 587-602.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. Journal of Educational Measurement, 19(2), 149-154.