SMALL SAMPLE ESTIMATION OF ITEM PARAMETERS IN ITEM RESPONSE THEORY MODELS USING OPERATIONAL DATA

MASTER OF EDUCATION (TESTING, MEASUREMENT AND EVALUATION) THESIS

BY

TAMANDANI AUGUSTINE CHIKOKO

Submitted to the Department of Education Foundations, Faculty of Education, in partial fulfilment of the requirements for the degree of

Master of Education (Testing, Measurement and Evaluation)

University of Malawi

Chancellor College

November, 2013

DECLARATION

I, the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

TAMANDANI AUGUSTINE CHIKOKO

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature_____Date_____

Bob W. Chulu, PhD (Senior lecturer)

Main Supervisor

Signature_____Date_____

Richard Nyirongo, PhD (Senior lecturer)

Head of Department

DEDICATION

I dedicate this work to my parents Mr and Mrs Chikoko who defied all odds to educate us all their children. May the Good Lord richly bless you.

ACKNOWLEDGEMENTS

The successful completion of this thesis would not have been possible without the support, guidance and assistance of many people, amongst others, the following need special mention:

Special thanks goes to my Thesis Supervisor, Dr. Bob W Chulu for his unwavering, intellectual, moral and other forms of support and understanding that he rendered throughout the course of this study. Special thanks to my parents Mr. and Mrs. Chikoko for giving me both moral and financial support throughout this study. I am also highly indebted to my entire family for their moral support and encouragement. I would like to sincerely thank my research assistant, Mr. P. Luciano who spent a number of hours on the computer entering the data.

ABSTRACT

Item Response Theory (IRT) models have been widely used to analyse test data and develop IRT-based tests. An important requirement in applying IRT models is the stability and accuracy of model parameters. Substantial research work has been undertaken in the past to study the effect of sample size on the estimation of IRT model parameters using simulations. One of the limitations of using pure simulations to study the effect of sample size on IRT item parameter estimation is that the model assumptions are strictly met, which is seldom true for operational test data. However, data from operational tests do not normally strictly meet the model assumptions. It was therefore in the interest of this study to use real data in comparing item parameter estimates from different samples sizes so that the possible minimum sample size could be determined for application in IRT dichotomous models. The study compared three sample sizes of: 250,500 and 1000 obtained by administering a 60 item multiple choice test to 1750 MSCE students across Zomba City. The analysis was done using ANOVA in SPSS. At 95% confidence interval the results showed that the item parameter estimates obtained from the three independent samples were statistically the same. This lead to the conclusion that a sample of size 250 can be employed in IRT's 2PLM and 1PLM to obtain item parameter estimates that are statistically equivalent to those from larger samples.

TABLE OF CONTENTS

TABLE **PAGE**

DECLARATION……i

CERTIFICATE OF APPROVAL ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

ABSTRACT…….....v

TABLE OF CONTENTS……...……vi

LIST OF FIGURES ix

LIST OF TABLES x

LIST OF ACRONYMS xi

CHAPTER ONE: INTRODUCTION TO THE PROBLEM

1.1 Background to the problem……1

1.2 Statement of the problem 2

1.3 Purpose of the study 3

1.4 Research questions 3

1.5 Significance of the study 4

1.6 Operation definition 4

CHAPTER TWO : REVIEW OF RELATED LITERATURE

2.0 Chapter overview 6

2.1 Brieth introduction to item response theory models 6

2.1.1 IRT dichotomous models 7

2.1.2 the one parameter logistic models 8

2.1.3 the two parameter logistic models 9

2.1.4 the three parameter logistic models 10

2.2 Irt parameter estimastion methods 10

2.2.1 Joint maximum likelihood method 11
2.2.2 Marginal maximum likelihood method 13
2.2.3 Bayesian methods 15
2.3 Invariance property of item response theory 16
2.4 Assumptions of Irt 17
2.4.1 Unidimensionality assumption 17
2.4.2 Local independence assumption 19
2.5 Sample size requirements for irt models 19
2.6 Sample size versus estimation methods 20
2.7 Samples versus modified models 21
2.8 Sample size versus optimal examinees 27
2.9 Adequate sample size……..……30

CHAPTER THREE : RESEARCH DESIGN AND METHODOLOGY OF THE STUDY

3.0 Chapter overview 33
3.1 Design of the study 33
3.2 Sampling 33
3.2.1 Selection of schools 34
3.2.2 Selection of examinees random samples 34
3.3 Data instruments 35
3.4 Validity and reliability of the questionaire` 35
3.5 Data collection 36
3.6 Data analysis 36
3.7 Ethical considerations 36

CHAPER FOUR: RESULTS OF THE STUDY

4.0 Chapter overview 38
4.1 Preliminary results of the study 38
4.1.2 Model data fit analysis 39
4.1.3 Model assumptions of unidimensionality and local independence 39

4.1.5 Graphic analysis of model data fit standardized residuals 40
4.1.6 Frequency distribution for standardized residuals for real and simulated data 40
4.1.7 Standardised item residual plot 42
4.1.8 Chi-square statistics 44
4.1.9 Predicted score distribution 44
4.1.9.1 Preliminary results summary 46
4.2 Main findings of the study 46
4.2.1 Item parameter estimates 46
4.2.2 Graphical comparison of item parameter estimate 50
4.2.3 Graphical assessment of item discrimination parameter estimates 50
4.2.4 Graphical assessment of item difficulty parameter estimates 53
4.3 Comparison of item parameters in anova 50
4.3.1 Anova results for 2pl item difficulty parameter 56
4.3.2 Anova results for item dscrimination parameter estimates 59
4.3.3 Anova results for person ability parameter estimates 61

CHAPTER FIVE: DISCUSSION, CONCLUNSIONS, IMPLICATIONS AND RECOMMENDATIONS

5.0 Chapter Over View 65
5.1 Discussion 65
5.2 Relationship Of The Findings To Prior Research 67
5.3 Implications For Practice And Policy 68
5.4 Limitation Of The Current Study 70
5.5 Recommendations 70
References 71
Appendices 79

LIST OF FIGURES

Figure1: 3 PLM ICC.……...8

Figure2: Scree plot for 1000 sample size data set……..40

Figure3: SR distribution for 1000 sample size data set 1pl model……41

Figure4: SR distribution for 1000sample size data set 2pl model……41

Figure5: SR distribution for 250 sample size data set 1pl model……..42

Figure6: SR distribution for 250 sample size data set 2pl model……..42

Figure7: SR distribution for 500 sample size data set 1pl model……..42

Figure8: SR distribution for 500 sample size data set 2pl model……..42

Figure9: SRs for 1000 sample size data set 1pl model……..43

Figure10: SRs for 1000 sample size data set 2pl model……43

Figure11: Score cumulative distribution for 1000 sample size data set 1pl model……...45

Figure12: score cumulative distribution for 1000 sample size data set 2pl model……45

Figure13: Item discrimination parameter estimates from 2pl model item 1 to 10…….50

Figure15: Item discrimination parameter estimates from 2pl model item 11 to 20……...51

Figure16: Item discrimination parameter estimates from 2pl model item 21 to 30……...52

Figure17: Item discrimination parameter estimates from 2pl model item 31 to 40…..……..52

Figure18: Item discrimination parameter estimates from 2pl model item 41 to 50 ……..52

Figure19: Item difficulty parameter estimates from 2pl model item 1 to 10……..…53

Figure20: Item difficulty parameter estimates from 2pl model item 11 to 20…….…….....54

Figure21: Item difficulty parameter estimates from 2pl model item 21 to 30……...……54

Figure22: Item difficulty parameter estimates from 2pl model item31 to 40…...…….55

Figure23: Item difficulty parameter estimates from 2pl model item 41 to 50…….....55

Figure24: Item difficulty parameter estimates from 2pl model item 51 to 60……..……56

LIST OF TABLES

Table 4.1: Item Difficulty Parameter Estimates for Sample Size 250 47
Table4.2: Item Difficulty Parameter Estimates for Sample Size 500 47
Table4.3: Item Difficulty Parameter Estimates for Sample Size 100 48
Table4.4: Item Discrimination Parameter Estimates for Sample Size 250 48
Table4.5: Item Discrimination Parameter Estimates for Sample Size 500 49
Table4.6: Item Discrimination Parameter Estimates for Sample Size 1000 49
Table4.7: Descriptives for difficulty parameter estimates 57
Table4.8: ANOVA Statistics for difficulty parameter estimates……..... 58
Table4.9: Descriptive Statistics for Item Discrimination Parameter Estimates 59
Table4.10: ANOVA Statistics for Discrimination Parameter Estimates 60
Table4.11: Descriptive Statistics for Examinee Ability Parameter Estimates 62
Table4.12: ANOVA Examinee Ability Parameter Estimates 62

LIST OF APPENDICES

APPENDIX A: SRs for 1plm for sample of 250……...79

APPENDIX A: SRs for 1plm for sample of 500……...85

APPENDIX A: SRs for 1plm for sample of 1000…….....90

APPENDIX A: SRs for 2plm for sample of 250 ……..95

APPENDIX A: SRs for 1plm for sample of 500…….....104

APPENDIX A: SRs for 1plm for sample of 1000……...110

APPENDIXC: Instrument for the study……..101

APPENDIXD: Letter to the executive director of (MANEB) ……123

APPENDIXE: Letter to the South East Education Division Manager……124

LIST OF ACRONYMS AND ABBREVIATIONS

ANOVA	Analysis of Variance
CDSS	Community Day Secondary School
CFA	Confirmatory Factor Analysis
CTT	Classical Test Theory
EFA	Exploratory Factor Analysis
1PLM	One Parameter Logistic Model
2PLM	Two Parameter Logistic Model
3PLM	Three One Parameter Logistic Model
ICC	Item Characteristic Curve
IRF	Item Response Function
IRT	Item Response Theory
MSCE	Malawi School Certificate National Examination
MANEB	Malawi National Examination Board
PCA	Principal Component Analysis
RMSEs	Root Mean Squared Errors
SRs	Standardized Residuals
SPSS	Statistical Package for Social Sciences
TOEFL	Test of English as a Foreign Language

CHAPTER ONE

INTRODUCTION TO THE PROBLEM

1.1 Background to the problem

In recent decades, item response theory (IRT) models have been growing in popularity. The common IRT dichotomous models include (Rasch, 1PL, 2PL and 3PL).These models are increasingly being used in assessment programs due to the following advantages: firstly, they provide away to model the probability of giving a correct answer on an item based on the underlying ability of the examinee and item parameters.

Secondly, they provide information on item level and the leading property of invariance which stipulates that values of IRT item parameters ought to be identical for separate groups of examinees and through different measurement conditions (Rupp & Zumbo, 2006).

Despite being promising and increasingly growing in application, Item Response Theory (IRT) has one major setback which poses as a limitation in its application in assessment, in that it requires large samples to obtain accurate person and item parameter estimates. The problem with larges sample sizes is that they are costly, difficult or undesirable to obtain and they presents test security through item exposure problems (Wainer & Eignor, 2000).

In attempt to address the problem of large sample sizes, studies have been undertaken to determine the minimum possible sample size that can be employed to obtain accurate person and item parameter estimates. As early as 1968, Lord suggested using test lengths of at least 50 items and sample sizes of at least 1,000 when using JML to estimate 3PL model parameters in order to control the sampling error of the discrimination parameter estimates. Ree and Jensen (1983) examined several combinations of calibration and equating sample sizes. They suggested a minimum sample size of 500, but recommended administering test items "to the largest samples available" (p. 145). Results from studies generally indicate that the magnitude of the variation between sample estimates decreases with increasing sample size. However, the majority of the studies focused on the use of simulated data. One of the limitations of using pure simulations to study the effect of sample size on IRT item parameter estimation is that the model assumptions are strictly met, which is seldomly true for operational test data. It was therefore in the interest of this study to use real data in comparing item parameter estimates from different samples sizes so that the possible minimum sample size could be determined for application in IRT dichotomous models.

1.2 Statement of the problem

The dichotomous IRT models are flexible and useful way to score assessment data. However, their uses are limited due to reliance on large samples. Effective methods to improve the accuracy of IRT parameter estimation could result in an expansion of the models' use into areas of assessment in which they are currently unsuitable due to sample size limitations.

However many studies in this area have relied on the use of simulated response data to evaluate the extent to which sample size affects the accuracy and stability of IRT models in estimating item parameters. For example, Hambleton and Cook (1983) simulated tests of 10, 20, and 80 items with sample sizes of 50, 200, and 1000 in order to determine the effect of sample size on the standard errors of ability estimation curves. Ree and Jensen (1983) examined several combinations of calibration and equating sample sizes. They suggested a minimum sample size of 500, but recommended administering test items "to the largest samples available" (p. 145).

1.3 Purpose of the study

The purpose of this study was to find out whether item parameter estimates across different independent samples sizes of persons in IRT dichotomous models are statistically comparable using real data.

1.4 Research questions

The questions which the study was concerned with were stated. Answers to each of these questions were sought through testing of the null hypothesis derived from each of the questions:

1. Which IRT model fits the data?
2. How comparable are the item difficulty parameter estimates from different samples?
3. How comparable are the item discrimination parameter estimates from different samples?

4. How comparable are the examinees' ability parameter estimates from different samples?
5. To what extent are the item and person parameters from different samples different?

1.5 Significance of the study

As it was envisaged, this study has determined and established the minimum sample size which could be employed when generating item parameter estimates in dichotomous IRT models. The equivalence of parameter estimates across different samples has also been determined based on IRT frame works. The findings of this research study have added to the empirical knowledge on the influence of sample size on item parameter estimates based on IRT theoretical framework. Secondly, these findings could be used to reduce pretesting costs, because smaller samples would be sufficient. The findings will help improve test security by reducing item exposure (fewer examinees need to see each item to estimate the item parameters accurately). Finally, practitioners could use the flexible 2PL model in situations where populations are small or where a smaller calibration sample is desired.

1.6 Operational definition of terms

Item Response Theory (IRT): Hambleton and Jones (1993) state that, "Item response theory is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test.

Three Parameter logistic model (3PLM): It is an IRT model with three parameters (a, b and c) parameters where "a" is the discrimination parameter, "b" is difficult and "c" the guessing parameter.

Two Parameter Logistic Model (2PLM): It is an IRT model with two parameters (a and b) parameters

One Parameter Logistic Model (1PLM): It is an IRT model with one parameter (a) parameters

ANOVA: Analysis of Variance

Dichotomous Items: These are items that are scored wrong or correct e.g. multiple choice questions.

Polytomous Items: A polytomous item is one that has more than two score categories

Principal Components Analysis (PCA): It is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

Classical Test Theory: It is a theory that describes test scores by introducing three notions; test score (i.e., observed score), true score, and error score. All together the equation is as follows: X (observed score) = T (true score) + E (error score). At any time there are two unknowns in the equation for the examinee, thus, some assumptions must be made. First, true scores and error scores are uncorrelated; second, the average error score in the population is zero, and third; error scores in parallel tests are uncorrelated.

CHAPTER TWO

REVIEW OF RELATED LITERATURE AND RESEARCH

2.0 Chapter overview

The literature chapter begins with a brief introduction to IRT models, the property of item parameter invariance, model assumptions, followed by a summary of some of the most common IRT item parameter estimation methods, sample size requirement and the effects of different models on parameter estimation with small samples.

2.1 Brief introduction to IRT models

Much has been written about the theoretical foundations, development, and application of IRT (Hambleton et al.. 1991; Yen & Fitzpatrick, 2006; de Ayala, 2009). The intent of this section is to provide a concise introduction to IRT dichotomous models and a brief description of its benefits and limitations.

2.1.1 IRT dichotomous models

At its core, IRT is a group of statistical models used to analyse assessment data. These models, which focus on individual items rather than intact assessments, employ nonlinear functions to relate the properties of an item (e.g., difficulty, discrimination) to the probability of an examinee providing a particular response (e.g., correct, incorrect). Mathematically, this can be defined as

$$P_I(\theta) = P_I(X_i) = x_i[\theta], [\delta_i] \qquad \text{Equation1}$$

This equation, or item response function (IRF), indicates that the probability of an examinee responding *xi* on item *Xi* depends on one or more examinee ability parameters, $\{\theta\}$,and one or more item parameters, $\{\delta_i\}$. This equation illustrates the primary benefits of IRT: Because the probability of a given response is conditional on both the item and examinee characteristics, estimates of item parameters are (examinee) sample independent, and person estimates are independent of items (Hambleton, Swaminathan, & Rogers, 1991; Yen & Fitzpatrick, 2006). IRFs are displayed graphically using item characteristic curves (Yen & Fitzpatrick, 2006).

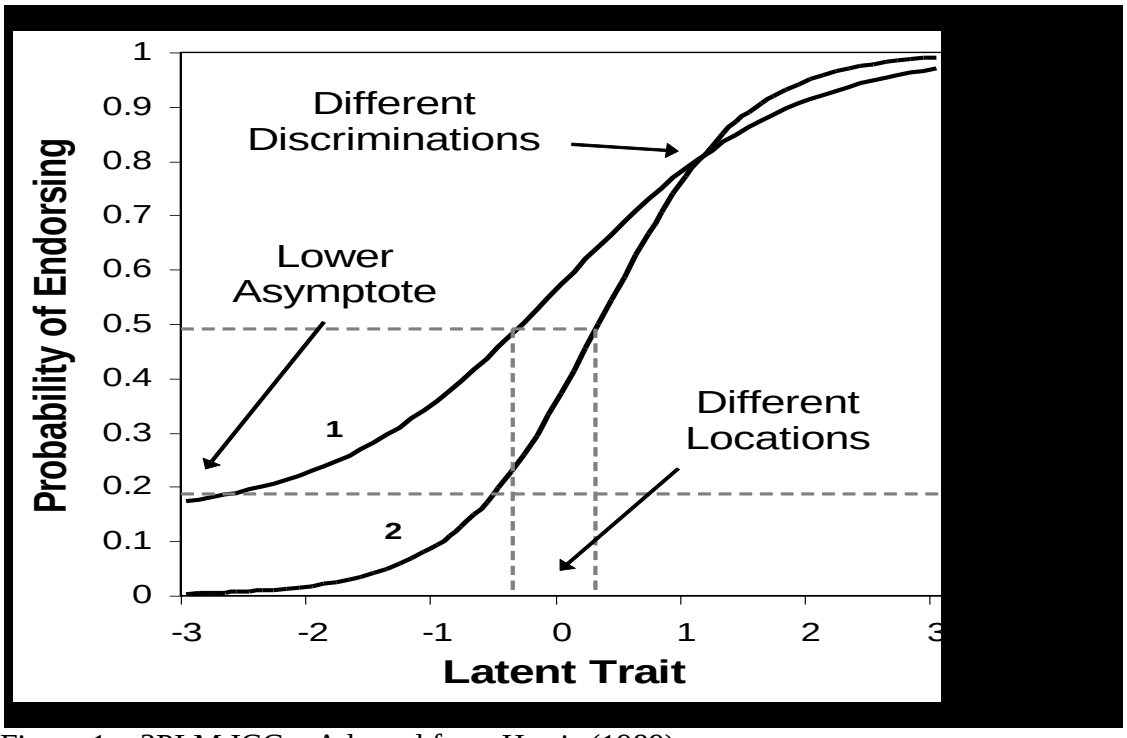


Figure 1: 3PLM ICC Adapted from Harris (1989).

Equation 1 is very general and does not specify that the item responses be either dichotomous or polytomous. This study focuses on an IRT model for dichotomous responses. The responses for dichotomous models are typically coded to either a zero (for an incorrect response) or one (for a correct response). Three of the most common models for dichotomous responses are discussed in more detail below.

2.1.2 The one-parameter logistic model

Equation 2

$$P(x_i = 1 | \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

This equation indicates that the probability of a correct response is dependent on the ability of the examinee (θ) and the item parameter *bi*, which is commonly referred to as item difficulty. Mathematically, the item difficulty corresponds to the ability level at the point of inflection of the ICC. Thought of another way, an examinee whose ability is equal to the item difficulty will have equal probabilities of getting the item correct or incorrect. When using the 1PL model, the shape of the ICCs is the same for all items; the ICCs merely shift up or down the θ scale depending on the item difficulty value (Hambleton, Swaminathan, & Rogers, 1991).

2.1.3 The two-parameter logistic model

An extension of the 1PL model is the two-parameter logistic (2PL) model

$$P(x_i = 1 | \theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Equation 3

This model is similar to the 1PL model but adds the additional item parameter, *ai*. The item parameter *ai* is commonly referred to as the item discrimination parameter and is a measure of the slope of the ICC at its point of inflection. Conceptually, item discrimination is an indication of the strength of the relationship between the item response and ability (Yen & Fitzpatrick, 2006). The constant *D* is often set to a value of 1.7 in order to make the model similar to the normal ogive function (Hambleton, Swaminathan, & Rogers, 1991). However, *D*'s value is a matter of individual preference (Yen & Fitzpatrick, 2006) and is not necessary.

For the 1PL and 2PL models it is a tacit assumption that as examinee ability levels become very low (approaching negative infinity), the probability of a correct response approaches zero. For many assessments, however, this may not be appropriate. For example, on multiple-choice assessments a low-ability examinee may get an item correct simply by guessing. The 3PL model allows for this possibility through the inclusion of a guessing parameter (sometimes referred to as a pseudo-guessing parameter). Therefore, extending the 2PL model results in the three-parameter logistic (3PL) model:

2.1.4 The three parameter logistic model

$$P\left(x_i = 1 \middle| \theta\right) = C_i + (1 - C_i)\frac{1}{1+e^{-Da_i(\theta - b_i)}} \qquad \text{Equation4}$$

Where *ai* and *bi* are defined above and *ci* is the pseudo-guessing parameter. Conceptually, the guessing parameter is the probability of a very low ability examinee getting an item correct. Mathematically; the guessing parameter is the value of the lower asymptote of the ICC.

2.2 IRT item parameter estimation methods

As stated above, the purpose of this study was to find out whether item parameter estimates across different independent samples sizes of persons based on IRT dichotomous models are statistically comparable using real operational data. With this in mind, it was useful to understand the traditional ways in which item parameter estimates are calculated. Three of the most common item parameter estimation methods are

summarized below: joint maximum likelihood (JML), marginal maximum likelihood (MML), and Bayesian estimation.

Although these are not the only item parameter estimation methods in use today, other methods tend to be less frequently used (e.g., nonparametric estimation) or specific to only a small number of models (e.g., conditional maximum likelihood (CML)). For a comprehensive review of the many IRT estimation methodologies refer to Baker and Kim (2004).

2.2.1 Joint maximum likelihood method

The solutions to the equations discussed in the previous section are complicated by the fact that in real testing situations parameters for both the items and the examinee abilities are unknown. JML addresses this problem by solving for both sets of parameters simultaneously. Let U be an $N \times n$ matrix consisting of dichotomously scored assessment results (1 = correct, 0 = incorrect) for an assessment that is n items long and administered to N examinees. Item responses are denoted U_{ij}, where i indicates the item, $i = 1, \ldots, n$, and j indicates the examinee, $j = 1, \ldots, N$. Let θ be a vector of ability parameters ($\theta_1, \ldots, \theta_j, \ldots, \theta_N$). Also, let $P_i(\theta_j)$ equal the probability of a person with ability $\theta_j$getting item i correct, and let $Q_i(\theta_i)$ equal 1-$P_i(\theta_j)$. Therefore, the probability of the observed results matrix, U, given the abilities of the examinees, θ, can be described by the following likelihood function:

$$L = Pro\left(U \mid \theta\right) = \prod_{j=1}^{N} \prod_{i=1}^{n} p_i^{uij}\left(\theta_i\right) Q_i^{1-uji}\left(\theta_j\right) \qquad \text{Equation 5}$$

Taking the natural log of Equation 5 yields

$$\ln L = \sum_{j=1}^{N} \sum_{j=1}^{N} \left[\quad u_{ij} \ln P_i((\theta_j) + (1 - u_{ij}) \ln Q_i(\theta_j) \right] \qquad \text{Equation 6}$$

The likelihood equation for a given parameter of interest, λ, is obtained by setting the first derivative of Equation 6, with respect to λ, equal to zero:

$$\frac{\Delta \ln L}{\Delta \lambda} = \sum \frac{p_i(\theta_j) - C_i(\theta_j)}{p_i(\theta_i) Q_i \theta_j \ \Delta \lambda} \left[u_{ij} - P_i(\theta_j) \right] = 0 \qquad Equation\ 7$$

For the parameters of the 3PL model, θj, ai, bi, and ci, Equation 7 can be rewritten as

$$\sum_{j=1}^{n} \frac{[p_i(\theta_j) - C_i]}{1 - c_1 p_i(\theta_j)} \left[u_{ij} - P_i(\theta_j) \right] = 0 \qquad \text{Equation 8}$$

For θj, $\frac{1}{1-c_1} \sum_{j=1}^{N} \frac{[\theta_j - b_i][p_i(\theta_j) - C_i]}{p_i(\theta_j)} \left[u_{ij} - P_i(\theta_j) \right] = 0$ Equation 9

For a_i,

$$\frac{\boldsymbol{a_i}}{\mathbf{1} - \boldsymbol{c_1}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{N}} \frac{\left[\boldsymbol{p_i(\theta_j)} - \boldsymbol{C_i} \right]}{\boldsymbol{p_i(\theta_j)}} \left[\boldsymbol{u_{ij}} - \boldsymbol{P_i(\theta_j)} \right] = \mathbf{0} \qquad \text{Equation 10}$$

For b_i, $\frac{1}{1-c_1} \sum_{\square=1}^{N} \frac{1}{P_i(\theta_j)} \left[u_{ij} - P_i(\theta_j) \right] = 0$ Equation 11

Equations 8-11 are solved using an iterative procedure with four steps:

Step 1. In the first step, person ability estimates are treated as fixed, set to an initial value, usually based on the examinee's raw score, and estimates are calculated for the item parameters.

Step 2. In the second step, the newly estimated item parameters are treated as fixed, and estimates are calculated for examinee abilities.

Step 3. In the third step of the estimation process, the difficulty and ability scales are set.

Step 4. New estimates are calculated for the item parameters while treating the newly estimated and re-centered person ability estimates as fixed.

Steps 2 through 4 are repeated until the change in parameter estimates between iterations becomes smaller than some fixed threshold known as a convergence criterion.

The second estimation procedure that will be discussed, MML, separates the estimation of item parameters from that of examinee abilities.

2.2.2 Marginal maximum likelihood

In JML both the item and examinee parameters are treated as fixed effects. Thus, as the number of examinees increases so do the number of parameters that need to be estimated. MML takes a different approach in that it treats examinees as random effects. The following description of the estimation process is a summary of a more detailed derivation by Baker (1987). It is assumed that the θ parameters are a random sample from an overarching normal distribution (or some other empirical or user-defined distribution),

(θ). As before, assume that the assessment is *n* dichotomous items in length. Let *s* equal the number of distinct response patterns and let *l* be the label of a specific response pattern such that $l = 1, 2, \ldots, s$. Therefore, the data matrix *U* is an $s \times n$ matrix consisting of one row for each of the *s* unique response vectors and one column for each of the *n* items. Therefore, the probability of an examinee with ability θ having the response vector u_l is

$$L_l = Prob(u = u_l) = \int_{-\infty}^{\infty} P\left(u = u_l \,|\, \theta\right) g(\theta). \qquad Equation\ 12$$

The integral is approximated by summing the estimated value of the probability function at *q* quadrature points, where *Xk* ($k = 1, \ldots, q$) is a specific quadrature point and *A(Xk)* is the quadrature weight of point *Xk*. Therefore, Equation 12 above can be approximated as follows

$$L_l = \sum_{K=1}^{q} \left(u = u_l \,|\, X_k\right) A(X_K) \qquad Equation\ 13$$

Equation 13 is used along with the expectation maximization (EM) algorithm (Bock & Aitkin, 1981) to obtain parameter estimates. In the first (E) step of the algorithm, initial item parameter estimates are used to obtain the expected number of examinees whose θ values correspond with the level of the quadrature point, $\overline{N_{\square}}$, and the expected number of correct responses to item *i* at that level, $\overline{r_{ik}}$. These values are estimated using the following equations:

$$\overline{N_k} = \frac{\sum_{l=1}^{s} r_l L_l(X_K) A(X_K)}{\sum_{k=1}^{q} L_l(X_K) A(X_K)}$$ Equation 14

$$\overline{r_{ik}} = \frac{\sum_{l=1}^{s} r_i u_{li} L_l(X_K) A(X_K)}{\sum_{k=1}^{q} L_l(X_K) A(X_K)}$$ Equation15

Where u_{li} is the response to item *i* within pattern *l*, and *Ll*(*Xk*) is the relative density at $\theta = Xk$.

In the second (M) step of the algorithm, $\overline{N_k}$ and $\overline{r_{ik}}$ are treated as observed data and used to obtain improved estimates of item parameters using the following equations:

$$\sum_{k=1}^{q} [\overline{r_{ik}} - \overline{N_k} P_i(X_k)] = 0$$ Equation 16

$$\sum_{k=1}^{q} \left[\overline{r_{ik}} - \overline{N_k} P_J(X_k)\right] X_k = 0$$ Equation17

2.2.3 Bayesian methods

Generally speaking, IRT Bayesian methods are modifications of either JML or MML estimation where a priori assumptions are made about the distribution of item parameters. These assumptions can be applied either formally or informally. For example the LOGIST software program (Wingersky, Barton, & Lord, 1982) uses JML along with an informal method of specifying the item parameter distributions by placing upper and lower limits on the *a* and *c* parameters (Mislevy & Stocking, 1989). In formal

applications of Bayesian methods, a prior distribution is specified and multiplied by the likelihood function to produce a posterior distribution from which parameter estimates are obtained (Baker, 1987). The BILOG software program (Mislevy & Bock, 1997) is based on MML estimation, but by default uses Bayesian methods of estimation for certain item parameters. For the 3PL model, discrimination parameters are assumed to follow a log-normal distribution, the difficulty parameters are assumed to follow a normal distribution, and the guessing parameter is assumed to follow a beta distribution. The specific parameters describing these prior distributions (i.e., hyper parameters) are either specified by the user or estimated from the data (Mislevy & Stocking, 1989).

2.3 Invariance property of item response theory

Tenants of IRT put forward the property of invariance possessed by parameter estimates, advocating that such estimates, are obtained free of context and can be deemed truly characteristic of their object, by opposition to the context-bound estimates in CTT. “Invariance” often means that values of IRT item parameters ought to be identical for separate groups of examinees and through different measurement conditions (Rupp & Zumbo, 2006).What is invariance? Like most authors on the same topic, Hambleton et al. (1991) stress the importance of this concept as a distinctive asset of IRT: The importance of the property of invariance of item and ability parameters cannot be overstated. This property is the cornerstone of item response theory and makes possible such important applications as equating, item banking, investigation of item bias, and adaptive testing.

On the one hand, “invariance” means equality: “If invariance holds the parameters obtained should be identical” (Hambleton et al., 1991, p. 20; Rupp & Zumbo, 2006, p. 64). On the other hand, a less stringent form of correspondence, e.g. linear equivalence, is admitted as a demonstration of invariance: two sets of parameters are said to be mutually “invariant” if they may be linearly transformed one into the other (Hambleton et al.. 1991; Rupp & Zumbo, 2006; Stocking and Lord, 1983).This second meaning of “invariance”, also named “congruence”, is akin to the notion of (linear) correlation, to the point that values of Pearson’s correlation coefficients are taken as conclusive indications of invariance (Fan, (1998), with a threshold value of $r = 0.90$ being proposed.

2.4 Assumptions of IRT

There are two assumptions underlying the model of IRT. These include *Unidimensionality* and *local independence* (Hambleton et al., 1991). These assumptions should be met in order to correctly fit data to a model.

2.4.1 Unidimensionality assumption

The assumption of unidimensionality affirms that only one type of ability can be measured by a group of test scores (Hambleton et al., 1991). This is not to say that other abilities cannot affect a test (i.e., levels of motivation and test anxiety), but that there should be a dominant factor which is sufficiently measured by the test (i.e., attachment; Hambleton et al., 1991). This assumption is sometimes difficult to meet because of “other” abilities, including cognitive and personality factors that can influence test performance (Hambleton & Swaminathan, 1985). In all, this assumption specifies the

importance of the evaluation through test scores of only one type of ability (Hambleton et al., 1991). Yet in reality, no scale in practice will ever be perfectly unidimensional (Harvey, 1999). As noted, the assumption of Unidimensionality is difficult to meet. Other factors including test motivation, cognitive skills, test anxiety, and test sophistication can influence the amount of abilities brought to a test. As such, these factors can influence the items and the predictability of the main ability in which the researcher may have wanted to study. For that reason, the construct must be well defined and validity evidence must be gathered to ensure that the test measures what it claims to (Hambleton, 1993).

There are a few of approaches which demonstrate that the assumption of Unidimensionality has been met. The first approach is to select a model and then fit the items to the chosen model. The second approach is to define the domains in which the researcher is interested in and then choose a model to fit the test. Items are pre-selected and factor analysis (i.e., measuring the variance in unobservable constructs) can be conducted to make sure that the items fit the dominant ability (Hambleton & Swaminathan, 1985). This is also called confirmatory factor analysis (CFA). Conversely, the main idea behind Exploratory Factor Analysis (EFA) is to investigate possible factors. Since it would be difficult to perfectly meet the assumption of Unidimensionality, some researchers contend that the main factor must make up at least 20% of the variance (Scherbaum, 2006). Consequently, it is up to the researcher to determine which approach is better in terms of meeting the assumptions of Unidimensionality (e.g. Principal component Analysis (PCA), Exploratory Factor Analysis (EFA); (Hambleton & Swaminathan, 1985).

2.4.2 Local independence assumption

The second assumption, local independence, states that when abilities influencing the test are held constant, responses to any item are statistically independent. This means that each item is independent of one another (Hambleton et al., 1991). When unidimensionality is met, local independence is usually met as well. Yet, local independence can still be met if unidimensionality has not been satisfied (Scherbaum, 2006). As a result, the complete latent space, which describes the process of inferring from an observed test score, will contain the dominant ability (Hambleton & Swaminathan, 1985).

Local independence specifies that scores on each test item do not present clues to the answers of any other test items. Since both assumptions are quite similar in terms of the latent space, factor analysis methods can also be employed for the assumption of local independence because once unidimensionality is met; local independence is assumed to be met (Hambleton, 1993). Unlike CTT, the data must fit the model chosen; which also infers local independence has been met (Dodeem, 2004).

2.5 Sample size requirements for IRT models

Some studies have shown that different IRT dichotomous models require samples of different sizes and the sample size should increase as the number of parameters to be estimated by the model increase. It has been argued that the 3pl model will require largest sample than the other two unidimensional models that is 2pl and 1pl models respectively (Lord, 1968; Hullin, Lissak & Drasgon, 1982; Talley, 2006). However in another study

smaller samples of 200 and 500 examinees were used in attempt to determine the effect of sample size on the standard errors of item and person parameter estimates and they proved to be sufficient, as adequate precision could be obtained using the sample of 200 examinees (Hambleton & Cook, 1983). In all these studies recommendation have pointed at using as largest samples as possible and consistently pointing at the sample of 1000 examinees as minimum for 3pl model however the majority of these studies used simulated data.

2.6 Sample size versus estimation methods

Several studies have been conducted to examine the effect of parameter estimation methods on the accuracy and stability of the estimates across samples of varied sizes. Some of the methods that have been compared include the CTT based point biserial correlations, Joint maximum likelihood (JML), Marginal Maximum likelihood (MML), Bayesian estimation methods, estimation heuristic procedures and the non-parametric estimation methods (Patsula & Gessoroli, 1995). The findings from these studies showed that some of the estimation procedures could not produce accurate and stable results with smaller samples but the differences become smaller as the sample sizes increased (Patsula & Gessoroli, 1995).

When the joint maximum procedure was compared to estimation heuristic procedures (Urry, 1974) and CTT's point biserial correlations methods across samples of; 250,500, 750, 1000 and 2000 examinees, the sample estimates resulted in correlations very similar

to those obtained through Joint maximum likelihood estimation as the sample size increased but with smaller samples differences were significant with the joint maximum producing better estimates (Jensema, 1972; Ree ,1979). When Bayesian estimation procedure was compared to the methods above using samples of: 100, 200 and 400 examinees on ability and difficulty parameter estimates the Bayesian procedure resulted into higher correlations and lower mean squared differences than did the other methods (Swaminathan & Gifford, 1986; Mislevy & Bock, 1984). In another study the Joint maximum procedure was compared to non-parametric estimation methods across samples of sizes; 100,250,500 and 1000 examinees, the findings showed that the non-parametric estimation methods produced stable and accurate results on smaller samples than compared to the joint maximum procedure (Patsula & Gessoroli, 1995).In another study Maximum likelihood procedure out performed joint maximum likelihood procedures on small samples however they matched on large samples (Yoes, 1995).

2.7 Samples versus modified models

Another way researchers have approached the problem of obtaining accurate parameter estimates with smaller sample sizes is to employ simplified/modified IRT unidimensional models. This section summarizes several studies in which researchers used simplified/modified unidimensional IRT models to analyze data. In 1983, Lord argued that when sample sizes are small, simple IRT models may provide more accurate results than more complex models, even when the more complex models theoretically should provide a better fit to the data. He evaluated this claim using item parameters derived from 1pl, 2pl and 3pl models using data from 3,000 sixth-grade students who took a 50-item Metropolitan vocabulary test. He concluded that when sample sizes were less than

200 the 1PL model resulted in more accurate ability estimates than did the 2pl and 3pl model.

In addition to using simplified models, researchers have examined the impact of using modified models. Barnes and Wise (1991) evaluated the efficacy of a 1PL model with a fixed nonzero lower asymptote. In their simulation study they examined this modified 1PL model with *c* fixed at one of two levels, .20 and .25. These models were compared to the 1PL (c = .0) and 3PL models across three sample sizes (50, 100, and 200) and two test lengths (25 and 50 items). The simulated data were based on ability and difficulty parameters generated from a standard normal distribution in the range of -3 to 3, .5 <a < 2.0, discrimination parameters ranged from .50 to 2.0, and guessing parameters that ranged from.10 to .30. Correlations, RMSEs, and bias of both the ability and the difficulty parameter estimates were used to evaluate results. Additionally, the RMSEs of the ICCs were examined. Five replications were carried out per cell. The 3PL model had the greatest problems with convergence. Ability estimates obtained using the modified 1PL models tended to have higher correlations with the true parameters than did estimates obtained using the 1PL and 3PL models. The modified 1PL model with the lower asymptote fixed at .20 produced the most accurate recovery of the ICCs. The authors suggested that a modified 1PL model may be the best in small sample estimations.

Sireci (1992) also examined the utility of modified IRT models, but used real rather than simulated data. The data were obtained from four administrations of a national financial planning certification examination over four years. Sample sizes were 173, 149, 106, and 159 examinees. The primary goal of the study was to evaluate the stability of item parameters for 13 test items that were common across all four test forms. Five IRT models were compared: the 1PL, 2PL, 3PL, modified 1PL (c =.20), and the modified 2PL (c =.20). The fixed value of the discrimination parameter was chosen to be the reciprocal of the number of answer choices (i.e., 4) minus .05. Item parameter estimates were obtained using MULTILOG. None of the models exhibited item parameter stability over the four data sets. Therefore, the author concluded that none of the evaluated models was appropriate for these small data sets.

Parshall, Kromrey, and Chason (1996) compared regular and modified IRT models with respect to model-data fit and stability. Simulated data were generated from 3PL item parameters obtained from a 40-item ACT mathematics assessment. Examinee abilities were generated from a standard normal distribution. Six models were examined: 1PL, 2PL, 3PL, modified 2PL (the discrimination parameter was restricted using a strong prior distribution), and two different modified 3PL models (the discrimination parameter was restricted using a strong prior distribution and one model had a common guessing parameter, which was estimated from the data, but constrained to be equal for all items). Four sample sizes were examined (100, 250, 500, and 1000). One hundred replications were conducted for each experimental condition. The BILOG software program was used for all calibrations. Model data fit was evaluated using item and person residuals.

Stability was evaluated using the standard deviations of the discrimination and difficulty parameters, and the ICCs across replications. The 3PL and modified 3PL (restricted *a*) models had the smallest item and person residuals for most sample sizes. However, these models had the least stable difficulty estimates across replications; the most stable estimates were obtained using the 1PL and modified 2PL models. The most stable discrimination estimates were obtained from the models that constrained the discrimination parameters (i.e., the 1PL and modified models). Setiadi (1997) compared a modified 1PL model (c = .20) with the 1PL model (estimated using MML and several Bayesian variations) and the 3PL model. The 3PL model was estimated using the non-parametric TESTGRAF software program. The 1PL and modified 1PL models were estimated using BILOG. Item parameters for the simulation study were chosen from both real and hypothetical testing situations. Data were generated based on the 3PL model. The author examined two test lengths (30 and 60 items), three sample sizes (100, 200, and 500), two sets of item parameters (one taken from the Law School Aptitude Test and one created by the author with higher discrimination values) and two ability distributions (normal and uniform). One hundred replications were conducted for each condition. Results were evaluated using correlations, average errors, absolute bias, standard deviation of estimation errors, and RMSEs of item parameters. It was found that the modified 1PL model resulted in more accurate estimation of ability than did the other models when ability was normally distributed. For the uniformly distributed data, the modified 1PL model had the most accurate item parameter estimates.

Parshall, Kromrey, Chason, and Yi (1997) expanded on the earlier work of Parshall, Kromrey, and Chason (1996) by examining the efficacy of modified models in the presence of multidimensional data. As in the earlier study, six models were examined: 1PL, 2PL, 3PL, modified 2PL (the discrimination parameter was restricted using a strong prior distribution), and two different modified 3PL models (the discrimination parameter was restricted using a strong prior distribution one model had a common guessing parameter, which was estimated from the data, but constrained to be equal for all items). Simulated item parameters for an 80-item, 6-dimensional test were generated from archival assessment data. Examinee abilities were generated using independent standard normal distributions for each dimension. Four sample sizes were examined (100, 250, 500, and 1000), and one hundred replications were conducted for each experimental condition. Parameter estimates were obtained using BILOG. The authors used the same evaluative criteria as the earlier Parshall et al. study with the addition of the mean squared error of the expected response probabilities, the RMSE of the estimated number correct for each examinee, and the Spearman correlation of the estimated number correct score and the true number correct score. Results showed that the 2PL model provided the best fit to the data. However, with respect to estimation accuracy, the best results were obtained from the 3PL model and the modified 3PL model with restricted discrimination values. Clearly studies have shown that simplified/modified unidimensional models may be viable alternatives in situation where samples may be small.

However, these models are less helpful in situations where a relatively small sample is used to obtain item parameter estimates that are then treated as known and used to build

and/or administer a test based on the 3PL model (e.g., a large-scale CAT). Additionally, simpler models may result in worse estimates when the data fit a more complex model. For example, Hambleton and Cook (1983) found that for data generated with the 3PL model, the 3PL model resulted in more accurate rank-ordering of examinees than did the 2PL model.

In contrast to the studies described above, Stone, Weissman, and Lane (2005) compared competing IRT models with respect to the consistency of student proficiency classifications. That is, rather than examining the accuracy of ability estimates or scale scores, they evaluated the accuracy of classifications based on these estimates. This study used real data from 13,621 11th-grade students from a 1999 state mathematics assessment. The test consisted of 60 multiple-choice items. 1PL and 3PL models were fit using the MULTILOG software program. Using the bookmark standard-setting procedure, the score scale was divided into four categories: Below Basic, Basic, Proficient, and Advanced. A standard-setting panel used an ordered item booklet with the items ordered based on the 1PL model. The four performance categories were identified using the difficulties of three items. The same three items were used to compare student performance classifications based on the competing IRT models. Based on the two competing IRT models, students were classified into different performance categories about 10% of the time. In the same paper, the authors discussed the results of a simulation study based on the same data. That is, the item parameters were the 3PL estimates from the real data and abilities were generated from a standard normal distribution. With the simulated data, comparisons could be made between estimated and true performance classifications. When the 1PL model misclassified students, it tended to

underestimate their ability. However, under the 3PL model misclassifications were more equally balanced between under- and overestimation.

The studies described above provide comparisons of various competing IRT models. However, these models are less helpful in situations where a relatively small sample is used to obtain item parameter estimates that are then treated as known and used to build and/or administer a test based on the 3PL model (e.g., a large-scale CAT). Additionally, simpler models may result in worse estimates when the data fit a more complex model. For example, Hambleton and Cook (1983) found that for data generated with the 3PL model, the 3PL model resulted in more accurate rank-ordering of examinees than did the 2PL model.

2.8 Sample size versus optimal examinees

Several researchers have attempted to improve item estimates (or obtain equally good estimates using smaller sample sizes) by choosing examinees in such a way as to get the most accurate item estimates possible. Wingersky and Lord (1984) investigated the effect that changing the number of items, number of examinees, and the distribution of examinee abilities had on the accuracy of item parameter estimates using real data from a regular administration of the Test of English as a Foreign Language (TOEFL). They used the 3PLmodel LOGIST for estimation (Wingersky et. al., 1982), and either a rectangular distribution with 1,500 examinees and 45 items or bell-shaped distributions of examinee abilities of either1,500 or 6,000 examinees and either 45 or 90 items. They found that the standard errors of item parameter estimates became smaller as sample size increased, but

were not substantially impacted by increasing the number of items. Conversely, they found that the standard errors of examinee ability estimates decreased as the number of items was increased, but were not substantially impacted by increasing the examinees. Additionally, they found that rectangular distributions of examinee abilities gave smaller standard errors for the item parameter estimates than did the bell-shaped distributions, indicating that better item parameter estimates could be obtained if examinees were selected systematically based on their ability. This recommendation was supported by a simulation study by Hambleton and Cook (1983), who found that the rank ordering of examinees was more accurate when examinee abilities were generated using a uniform distribution rather than a standard normal distribution.

Stocking (1990) expanded on the work of Wingersky and Lord by evaluating which examinee abilities provide the most information for estimating item parameters for the 1PLand 2PL models, as well as the 3PL model used in Wingersky and Lord (1984). Stocking showed that for the 3PL model:

• Both low and high ability examinees provide little information for estimating item discrimination (as do those with abilities close to the optimal value for estimating item difficulty); the most informative examinees have abilities just above or just below the item difficulty.

• Examinees provide the most information for estimating item difficulty when their ability is equal to the item's difficulty parameter, but when the guessing parameter is

greater than zero the optimal ability level for estimating difficulty is greater than the difficulty parameter and depends on the item's discrimination and guessing parameters.

• Only examinees with very low abilities provide information for estimating guessing parameters. Thus, the examinee that provides the most information for estimating the difficulty parameter may be very different from the examinee who provides the most information for estimating the discrimination parameter, who also may be different from the examinee who is most useful for estimating the guessing parameter. Stocking concluded that selecting samples of examinees where ability was distributed either uniformly or bi-modally would serve as a good compromise for overall item parameter estimation accuracy.

From these studies it is evident that there exist several factors affecting the stability and accuracy of parameter estimates apart from the sample size and length of a test, these factors include: the model used to generate the parameters, the ability distribution of the population and estimation procedures. However in examining these factors majority of the studies used simulated data sets to generate parameters estimates, the problem with simulated data is that model assumptions are strictly meet which is hardly the case with real data which most of the time is messier.

2.9 Adequate sample size

Lord (1968) Lord calibrated the 3PL model on SAT data by an iterative technique that closely resembled Joint Maximum Likelihood estimation (JMLE), but did not include maximum likelihood estimation of the c parameter (it was estimated in an ad hoc non-parametric manner). After much difficulty, he was able to obtain convergence. In describing his difficulties, Lord commented that the sampling errors of the estimated discrimination parameters "seem to be excessive unless n > 50, perhaps, and N > 1000," where n is the number of items, and N is the number of examinees.

Hulin, Lissak, and Drasgow (1982). They referenced Lord (1968) as recommending a sample size of at least 1000 examinees and investigated sample sizes 200, 500, 1000, and 2000 with repeated simulation trials to get a better idea of the effect of sample size Like Lord (1968); they estimated parameters using a form of JMLE, as operationalized in the LOGIST computer program. In regard to recovery of the true item characteristic curves (ICC's), they reported average RSME values of about 0.03, 0.04, 0.05, and 0.06, for sample sizes of 2000, 1000, 500, and 200, respectively, for a test length of 60 items. Hulin et al. did not make a specific recommendation with respect to sample size, but others have referenced them as recommending 1000 examinees and 60 items (Refer to Baker, 1992, p. 106).

Unfortunately, the JMLE method has since been found to be inconsistent (not guaranteed to converge as sample size increases) (Little & Rubin, 1983). In spite of this, many

researchers have referenced both Lord (1968) and Hulin et al. (1982) as recommending a sample size of at least 1000 examinees for calibrating the 3PL model.

Mislevy (1986) applied MMLE with a sample size 1000. But Mislevy's paper was mostly theoretical in nature, and the estimation was conducted with only a single simulated dataset as a demonstration of the procedure without any significant conclusions about the adequacy of the estimation. Still, others have referenced Mislevy's article as support for the use of 1000 examinees and have interpreted his results as showing that the item parameters were accurately recovered(e.g., Harwell & Janosky, 1991).

Yen (1987), investigated MMLE (as implemented in BILOG) using a sample size of 1000. She reported that the RMSE for the difficulty and ability parameter estimators were approximately 0.15 and 0.10, respectively, for a 40-item test of moderate difficulty (similarly good results for other realistic settings were also reported), thus giving significant support to the use of 1000 examinees as an adequate sample size.

Gao and Chen (2005), looked at sample sizes of 100, 500, and 2000. For the case of 2000 examinees and 60 items (the most realistic in comparison with typical standardized tests), the RMSE was about 0.11 for a parameter estimation, 0.12 for b parameter estimation, and 0 (to the nearest hundredth) for c parameter estimation, with correlations between estimated and true values being 0.97, 1.00, and 1.00, respectively. These results certainly give strong support that a sample size of 2000 is more than what is needed.

These studies can both be interpreted as lending support to the adequacy of using a sample size of 1000 in calibrating the 3PL model. The combined results of all the 3PL studies, with more emphasis given to the MMLE results of Yen (1987), Hanson and Beguin (2002), Gao and Chen (2005), and Kim (2006), all seem to indicate that the use of 1000 examinees can be depended upon to give adequate parameter estimation results. However, majority of the studies both real and simulated concentrated on determining the adequate sample size for the 3pl model alone without due consideration of the other dichotomous models (1plm and 2plm). It has also been reported that most of these studies employed the JMLE method but Unfortunately, the JMLE method has since been found to be inconsistent (not guaranteed to converge as sample size increases) (Little & Rubin, 1983). In spite of this, many researchers have referenced both Lord (1968) and Hulin et al. (1982) as recommending a sample size of at least 1000 examinees for calibrating the 3PL model, Though the study had convergence problem which might have affected the accuracy of parameter estimation. It is therefore the lack of empirical research in the other dichotomous models concerning sample size that has compelled this study to take place. This study used Bayes prior information about item parameters to improve estimation convergence over MML and JML (e.g., Kim, 2007; Mislevy, 1986; Swaminathan & Gifford, 1986).

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

3.0 Chapter overview

This discusses the design of the study, the sampling procedure, data generation, instrument, procedure and technique that were used to analyse the data and ethical consideration that guided the research process.

3.1 Design of the study

The study employed single factor experimental design the one factor variable that was manipulated was sample size whilst the item parameters estimates were dependent variables. The sample size independent variable had three levels; 250, 500 and 1000 examinees.

3.2 Sampling

To examine the issues related to the effects of small sample sizes on IRT statistics, three examinee samples of varied sizes were implemented for the MSCE English language test data so that the behaviors of IRT statistics could be examined under different sample conditions as follows: 250,500 and 1000.This section also explains how the participating schools were selected from the total population of 40 schools.

3.2.1 Selection of schools

The participants were candidates for 2013 Malawi School Certificate Education selected randomly from eight of the forty: Boarding secondary Schools, Day secondary schools and Community secondary schools within Zomba. These schools were selected using simple random sampling technique with each of the 40 schools having an equal inclusion probability of $1/5$.

3.2.2 Selection of examinees random samples

Test scores were collected from 2000 examinees from these examinees three independent samples of varied sizes were created for the study as follows: 250, 500, and 1000. A sample Size of 1,000 is usually considered the minimum for use with the 3PL model. Therefore, sample size of 1000 examinees was included in order to serve as a benchmark for the smaller sample sizes. The sample sizes chosen here were similar to those used in other IRT simulation studies (Harwell, 1996).

From the 8 school in which the instrument was administered a pool of 2000 examinees was collected in order to create the three samples, systematic sampling (SYS) was used, to create the three examinees samples from the pool of 2000 at regular intervals. With population N = 2000, and multiples samples of sizes n_1 = 250, n_2 = 500 and n_3= 1000, every k^{th}unit is selected where the interval k was equal to N/n_i, the random start, r, was a single random number between 1 and k, inclusively. The units selected were then: r, r+k, r+2k... r+ (n-1) k. with systematic sampling each unit had an inclusion probability, π, equal to n/N.

The researcher chose to use systematic random sampling because it offers the following advantages:

Firstly, it is an alternative for simple random sampling (SRS) when there is no frame and it does not require auxiliary frame information. Secondly, it can result in a sample that is better dispersed, Babington, (1975). Thirdly, systematic procedure has a well-established theory and so estimates can be easily calculated.

3.3 Data generating instruments

The study used a MANEB 2009 English language MSCE paper for generating real data that was used in the study. The length of the test used was 60-items. This length was selected in order to provide a standard that is representative of assessments being used in the field. For example, Hambleton and Cook (1983) reported that "a test with 10 items is the shortest a test as is ever used in practice". Sixty items is generally believed to be adequate even for most 3PL model applications. Hulin, Lissak, and Drasgow (1982), used 10 item tests to save as a yard stick for the smaller item tests. The test length chosen here is also similar to test length used in other IRT simulation studies (Harwell, Stone, Hsu, & Kirisci, 1996).

3.4 Validity and reliability of the questionnaire`

The instrument was developed by Malawi Examinations Board MANEB and was piloted and used on a large population in 2009 therefore the instrument can be said to be psychometrically sound (reliable and valid). Assessment of Model assumptions for unidimensionality and local independence using CPA was conducted and results showed

that the mentioned requirements were met. This provided the evidence for construct validity which demands unidimensionality.

3.5 Data generation

The data was be generated by administering multiple choice test using 2009 Malawi Secondary Certificate Examination English Language Paper to 2000 form4 students from randomly selected eight secondary schools in Zomba, The participants were chosen on the understanding that they had covered the MSCE English syllabus since the data collection was done just a month before they sat for their 2013 Malawi Secondary Certificate of Education examinations.

3.6 Data analysis

The software BILOG MG VERSION 7.0 was used to generate IRT item parameters estimates. One way ANOVA was used for testing hypothesis on the item difficulty and discrimination parameter estimates in SPSS whilst Resid-plots program was used to assess model data fit by graphically analysing the standardised residual distributions the IRT model assumptions of unidimensionality and local independence were assessed using Principal Component Analysis in SPSS.

3.7 Ethical considerations

Ethical issues and standards were critically considered in this research project. According to Strenbert and Carpnter (1999) the aim of ethical considerations in research is to do well to the subjects of the study and avoid any harm. Therefore to meet the said standards the researcher negotiated access to schools from, South East Education Division

Manager, Head teachers of all participating schools and the Malawi National Examination Board (MANEB) refer to appendices (E, F, and G), and the examinees were given understandable explanation of the purpose of the study and the procedure to be followed. Participation was voluntary and they were free to withdraw from the study at any time.

CHAPTER FOUR

RESULTS OF THE STUDY

4.0 Chapter overview

This section presents the preliminary result and the main finding of the study. The preliminary results includes: Assessment of model assumptions of unidimensionality and local independence, the model data fit assessment which was conducted by analysing the standardised residual plots and distributions produced from 1pl and 2pl models and the Chi-square statistic. The section of the main findings contains the graphical and the statistical comparison of item and person parameter estimates

4.1 Preliminary results of the study

The purpose of this study was to find out whether item and person parameter estimates across independent samples of different sizes of examinees in IRT dichotomous models are statistically comparable using real data. This section presents results for the; Assessment model assumptions unidimensionality and Local independence, Generation of item parameters estimates in BILOG, Model- Data fit analysis with the data and item and person parameters estimates done using Resid-Plot program and SPSS.

4.1.2 Model data fit analysis

In this study, the data from three samples of real data was fitted to one, two and three parameter models .The samples were of different sizes: 250, 500 and 1000. The sample size was the independent variable while the item parameter estimates derived from these models were the dependent variable.

According to Box, G.E.P. (1979) A "model " is something we use to approximate reality for the purpose of making predictions, explaining data, etc. Strictly speaking, no model will fit the data perfectly but the question is, "how much model-data misfit is too much?"

In trying to make choice of which Model to use in this study the researcher employed the following assessment model- data fit techniques: Evaluation of model assumptions, assessment of residuals and standardised residuals plots, and the chi-square model fit statistic and the comparison of observed and simulated distribution.

4.1.3 Model assumptions of unidimensionality and local independence

In this study, *a scree plot* generated from Principal Components Analysis (PCA), was used to evaluate the dominance of the first factor. The figures below represent the scree plot for data from the sample of 1000 examinees. The rule of thumb requires that the first factor accounts for 20% of the variability in the data.

From the scree plot produced in the PCA from the sample of size 1000 it is evident that the first dominant factor exists in the data, this confirms the assumptions of unidimensionality and local independence.

Figure 2: Screeplot for 1000 sample size Data set

4.1.5 Graphic analysis of model data fit standardised residuals

Standardised residuals are the basis for the plot for each item, standardised residual distribution item fit plot, and score fit plot. It is calculated as follows:

$$SR_j = \frac{(\boldsymbol{O_j - E_j})}{\sqrt{\frac{\boldsymbol{E_j(1-E_j)}}{\boldsymbol{N_j}}}}$$

Where $O_j$is the observed proportion of correct answers for examinees in a score interval, $E_j$is the expected (model-based) proportion of correct answers in the same score interval, N_j is the number of examinees in the same score interval.

4.1.6 Frequency distribution for standardised residuals for real and simulated data

The standardised residual frequency distribution is based on all SRs (intervals ×Items) in the test excluding those with intervals with zero frequencies.

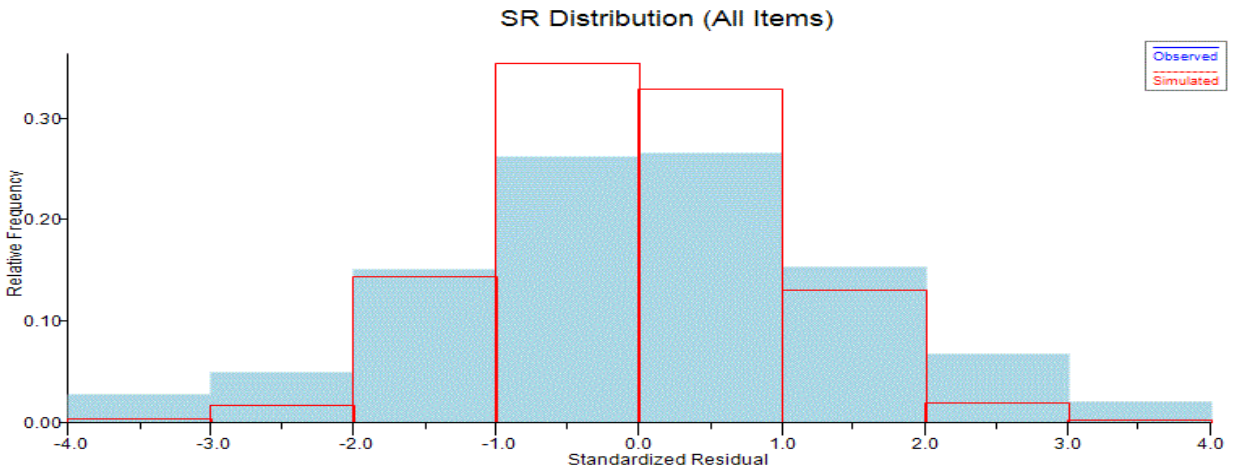


Figure 3: SR Distribution for 1000 sample size data set 1PL Model

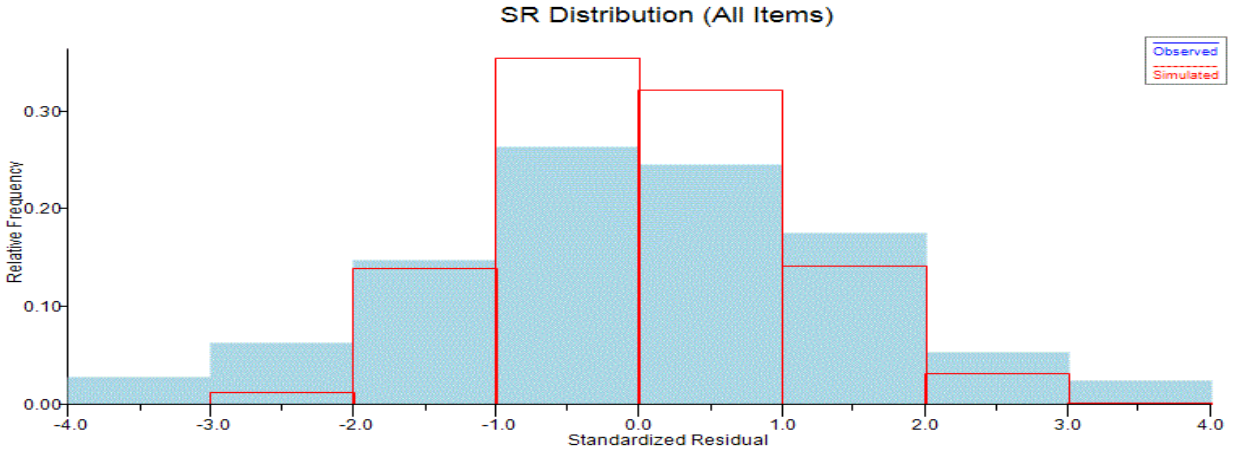


Figure 4: SR Distribution for 1000 sample size data set 2PL Model

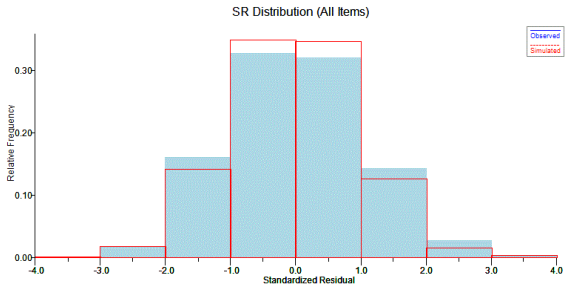


Figure 5: SR Distribution for 250 sample size data set 1PL Model

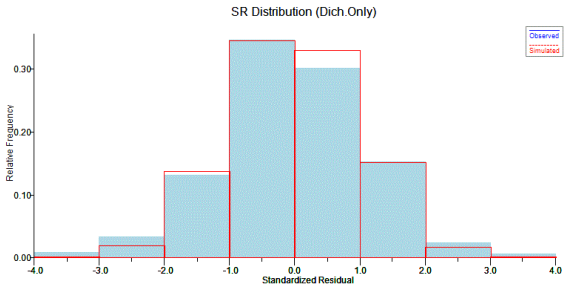


Figure 6: SR Distribution for 250 sample size data set 2PL Model

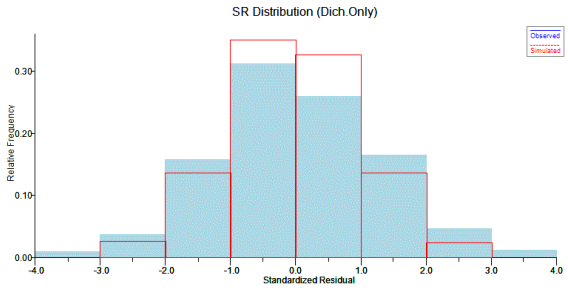


Figure 7: SR Distribution for 500 sample size data set 1PL Model

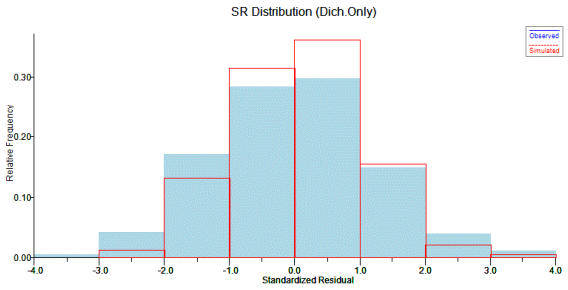


Figure 8: SR Distribution for 500 sample size data set 2PLModel

The Figures display the real and simulated distribution for the Sample Size 1000,250 AND 500from 1PL and 2PL models respectively. It is evident in both cases that both the real and simulated distribution are identical i.e., they all have a normal distribution in 2PL models unlike the situation in the 1PL model.

4.1.7 Standardised item residual plot

The standardised residual in each score interval is shown on the plot. If there are no Examinees in an interval, the standardised residual is not shown in the display.

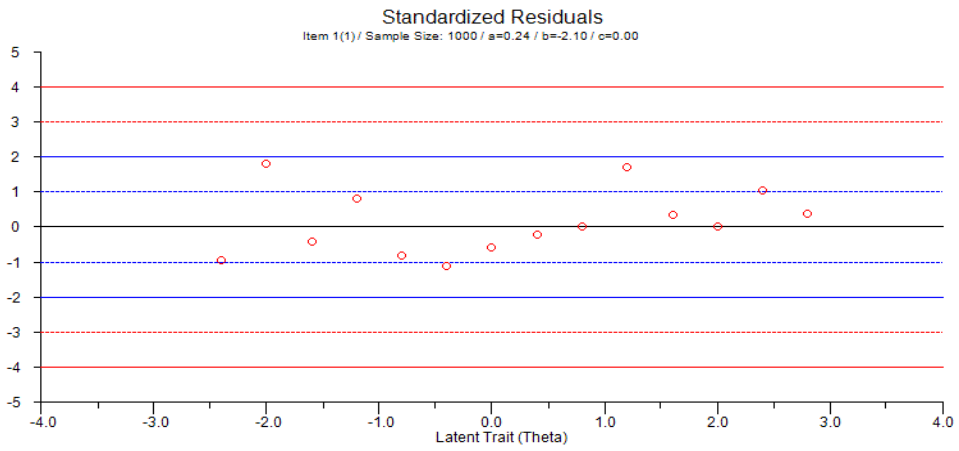


Figure 9: SRs for 1000 sample size data set 1P Model

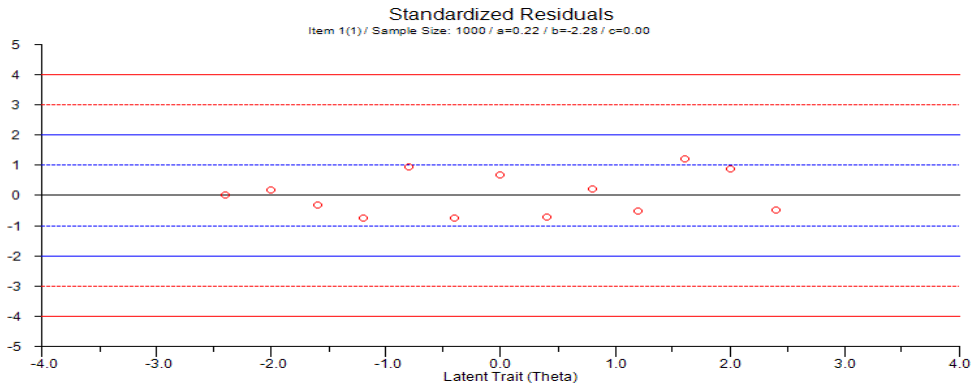


Figure 10: SRs for 1000 sample size data set 2P Model

Only one item standardised residual plot for item one is presented for both 2PL and 1PL models rests of the plots are given in the appendix. It is evident from the graphs that most residuals are homoscedastic (having equal standard deviation) for each item and follow an approximately standard normal distribution across all items of the test. The 2PL model gives standardised residual deviations smaller range of (-1 to +1) compared for 1PL model range of (-1 to +2). Therefore the researcher concluded that the 2PL model fits the Data better than 1PL model.

.

4.1.8 Chi-square statistics

This statistic from is reported in the Fit STAT table. For each item, it is calculated as follows:

$$x^2 = \sum_{j=1}^{K} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}$$

For the sample size 250, only 17% of the items were demonstrated misfit in both 1P and 2P model with alpha set at 0.05. At the sample level of significance 40% of the items demonstrated misfit when sample size 500 was fitted to 1P model and 28% when the same sample was fitted to 2PL. Therefore with the chi-square statistic, 2P model is 12% better than 1pl model.

4.1.9 Predicted score distribution

As was the case with standardised residuals real –simulated data distributions, in predicated score distributions, the actual test score distribution was compared with the distribution of predicated test score. When they are close, it is said that the best fitting IRT model closely recovers or predicts the actual test score distribution for the examinees that were administered the test. When they are not close, model fit can be questioned. It is a judgment as to how close the distributions need to be to establish model fit. Interpretation is enhanced by comparing the fit for more than one model to provide a basis for interpreting the results.

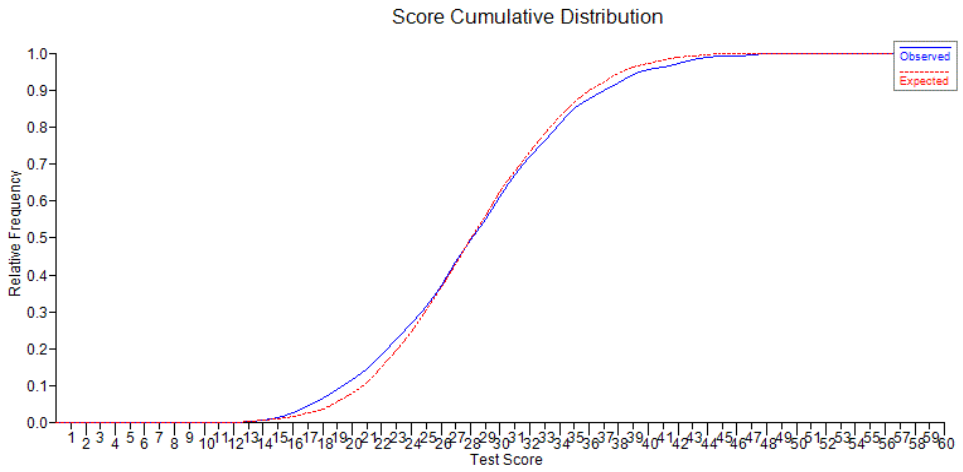


Figure 11: Score cumulative distribution for 1000 sample size data set 1P Model

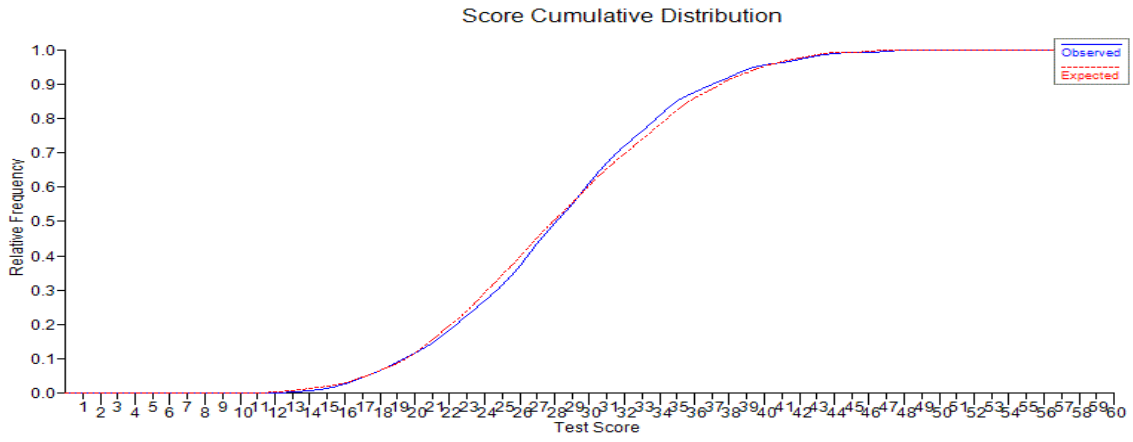


Figure 12: Score cumulative distribution for 1000 sample size data set 2P Model

In figures, 11 and 12 the 1PL and 2PL model respectively, the observed are closely fitting to the expected in the middle of the distribution however the test score distribution for 1PL model the observed distribution is slightly deviating expected distribution at both ends. Hence, the conclusion that 2PL model fits the data better than the 1PL model.

4.1.9.1 Preliminary results summary

In summary, after thorough assessment of unidimensionality and local independence of the data and analysis of the standardised residual plots and distributions, predicted score distribution and the Chi-square Fit statistic and other graphical item and test fit analysis included in the appendix. The researcher concluded that 2PL model fits the data better than the 1PL model. Therefore 2PL model were chosen for this study. The 3PL model was left out due to convergence problem with the BILOG .MG3 IRT program.

4.2.0 Main findings of the study

The section reports item parameters generated across the three samples, the graphical and statistical comparisons results obtained from item and person parameter estimates across the three samples. The parameters were generated in BILOG .MG3 IRT program, using the programs default setting in all cases.

4.2.1 Item parameter estimates

Upon choosing 2PL as the best model among three, the researcher generated item difficulty, discrimination and examinees ability estimates using the three examinees' samples. This section presents the item difficulty and discrimination parameters estimates which were compared to examine possible differences with respect to sample sizes

Table 1

Item Difficulty Parameter Estimates for Sample Size 250

item #		item#		item#		item#	
1	-0.789	16	-0.013	31	-1.077	46	-3.183
2	1.815	17	0.65	32	1.154	47	-3.017
3	1.394	18	261.054	33	-0.42	48	0.071
4	3.075	19	-1.296	34	0.991	49	-0.798
5	-0.879	20	2.395	35	-0.284	50	1.075
6	6.104	21	1.542	36	-0.407	51	0.278
7	3.653	22	0.517	37	-0.307	52	-1.102
8	0.271	23	-0.116	38	0.6	53	-0.607
9	0.763	24	0.117	39	-0.014	54	-0.633
10	0.169	25	0.82	40	-1.647	55	-1.179
11	-2.193	26	0.316	41	2.659	56	1.263
12	4.293	27	0.592	42	0.823	57	1.061
13	-0.563	28	1.567	43	-1.506	58	0.022
14	2.744	29	-0.584	44	0.593	59	-0.989
15	1.419	30	0.026	45	-1.223	60	2.387

Table 2

Item Difficulty Parameter Estimates for Sample Size 500

item #		item#		item#		item#	
1	-2.125	16	0.051	31	-0.963	46	-2.563
2	2.334	17	0.218	32	0.731	47	-2.131
3	0.552	18	3.556	33	-0.627	48	-0.095
4	2.55	19	-1.207	34	1.603	49	0.813
5	-0.819	20	3.899	35	-0.625	50	1.128
6	5.87	21	2.297	36	-0.513	51	0.19
7	3.595	22	1.205	37	-0.41	52	-1.246
8	0.662	23	-0.061	38	0.826	53	-1.189
9	0.532	24	0.25	39	0.02	54	-1.012
10	0.669	25	0.906	40	-1.362	55	-1.019
11	-2.614	26	0.266	41	3.069	56	1.414
12	3.999	27	0.962	42	1.725	57	1.155
13	-0.275	28	0.614	43	-0.684	58	0.074
14	3.999	29	-0.373	44	0.65	59	-0.93
15	1.342	30	0.086	45	-1.625	60	2.719

Table 3

Item Difficulty Parameter Estimates for Sample Size 500

item #		item#		item#		item#	
1	-2.104	16	0.52	31	-0.933	46	-2.639
2	1.818	17	0.162	32	0.748	47	-3.203
3	0.753	18	4.295	33	-0.63	48	0.394
4	2.733	19	-1.206	34	1.358	49	2.212
5	-0.658	20	6.014	35	-0.52	50	3.411
6	4.666	21	3.355	36	-0.312	51	0.203
7	7.027	22	1.395	37	-0.403	52	-1.175
8	0.68	23	-0.077	38	1.282	53	-1.457
9	0.535	24	0.17	39	0.166	54	-0.896
10	0.357	25	0.927	40	-1.035	55	-1.184
11	-2.833	26	0.308	41	3.613	56	1.178
12	5.429	27	1.648	42	3.906	57	0.933
13	-0.354	28	1.411	43	-0.825	58	0.012
14	3.763	29	-0.895	44	0.726	59	-1.015
15	1.771	30	0.293	45	-1.202	60	2.526

Table 4

Item Discrimination Parameter Estimates for Sample 250

item #		item#		item#		item#	
1	0.394	16	0.478	31	0.552	46	0.296
2	0.475	17	0.201	32	0.338	47	0.243
3	0.365	18	0.001	33	0.84	48	0.356
4	0.478	19	0.706	34	0.16	49	0.113
5	0.65	20	0.123	35	0.428	50	0.092
6	0.193	21	0.187	36	0.771	51	0.19
7	0.166	22	0.196	37	0.577	52	0.942
8	0.233	23	0.534	38	0.135	53	0.269
9	0.721	24	0.506	39	0.373	54	0.31
10	0.557	25	0.62	40	0.162	55	0.299
11	0.325	26	0.597	41	0.13	56	0.399
12	0.21	27	0.12	42	0.121	57	0.425
13	0.556	28	0.184	43	0.192	58	0.756
14	0.264	29	0.208	44	0.804	59	0.392
15	0.489	30	0.607	45	0.161	60	0.249

Table 5

Item Discrimination Parameter Estimates for Sample 500

item #		item#		item#		item#	
1	0.224	16	0.457	31	0.575	46	0.229
2	0.328	17	0.342	32	0.411	47	0.215
3	0.478	18	0.103	33	0.761	48	0.334
4	0.425	19	0.682	34	0.129	49	0.083
5	0.836	20	0.093	35	0.254	50	0.088
6	0.182	21	0.166	36	0.602	51	0.151
7	0.114	22	0.208	37	0.339	52	0.718
8	0.13	23	0.539	38	0.133	53	0.146
9	0.684	24	0.432	39	0.723	54	0.257
10	0.501	25	0.826	40	0.168	55	0.25
11	0.283	26	0.633	41	0.113	56	0.355
12	0.187	27	0.084	42	0.114	57	0.426
13	0.615	28	0.172	43	0.184	58	0.608
14	0.165	29	0.168	44	0.724	59	0.397
15	0.45	30	0.52	45	0.115	60	0.22

Table 6

Item Discrimination Parameter Estimates for Sample Size 1000

item #		item#		item#		item#	
1	0.241	16	0.286	31	0.695	46	0.249
2	0.42	17	0.286	32	0.399	47	0.147
3	0.478	18	0.066	33	0.786	48	0.221
4	0.465	19	0.568	34	0.105	49	0.053
5	1.074	20	0.063	35	0.28	50	0.049
6	0.223	21	0.118	36	0.728	51	0.118
7	0.079	22	0.122	37	0.448	52	0.818
8	0.09	23	0.432	38	0.078	53	0.122
9	0.717	24	0.567	39	0.685	54	0.298
10	0.489	25	0.75	40	0.139	55	0.203
11	0.274	26	0.476	41	0.066	56	0.409
12	0.148	27	0.052	42	0.064	57	0.497
13	0.65	28	0.119	43	0.139	58	0.617
14	0.163	29	0.134	44	0.623	59	0.358
15	0.314	30	0.439	45	0.097	60	0.173

4.2.2 Graphical comparison of item parameter estimate

This section reports comparison of item parameter estimates across the three samples by the way of graphing item *parameter vs. sample size.* For every item, the parameter estimates are compared across the three samples (250, 500, and 1000) in order to examine the trend and behaviour of item parameters across the samples (i.e. either increasing or decreasing).

4.2.3 Graphical assessment of item discrimination estimates

In the plots it is visible that the discrimination estimates are similar for each item across the samples. However in some cases the estimates are slightly larges in the 250 sample

than the other samples. In general there is an increasing trend as the sample size gets smaller. Therefore the researcher concluded that the item parameters across the samples were similar.

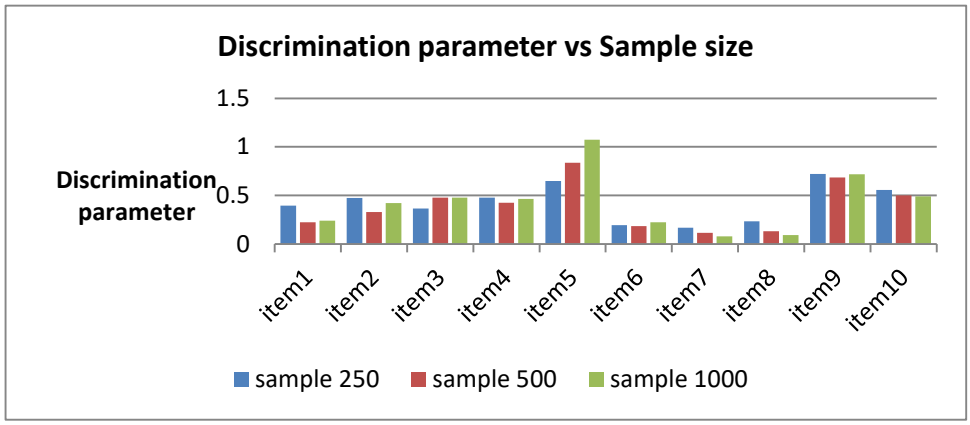


Figure 13: Item discrimination parameter estimates from 2plm

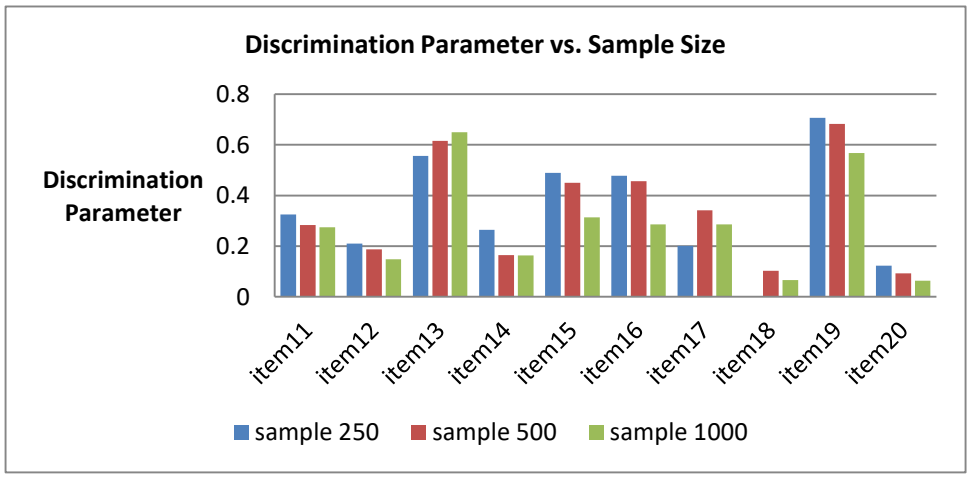


Figure 14: Item discrimination parameter estimates from 2plm

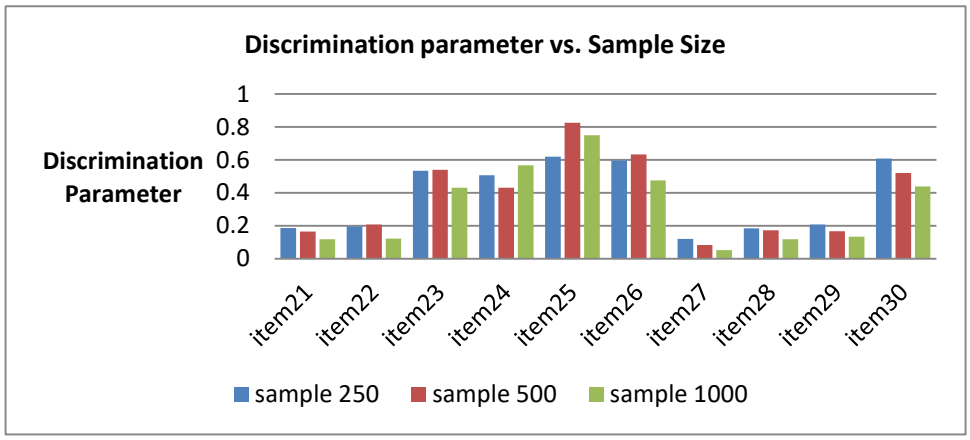


Figure 15: Item discrimination parameter estimates from 2plm

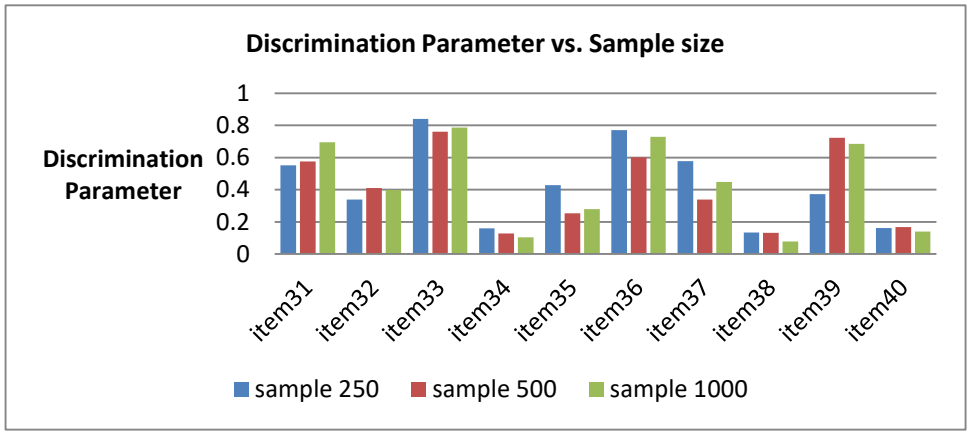


Figure 16: Item discrimination parameter estimates from 2plm

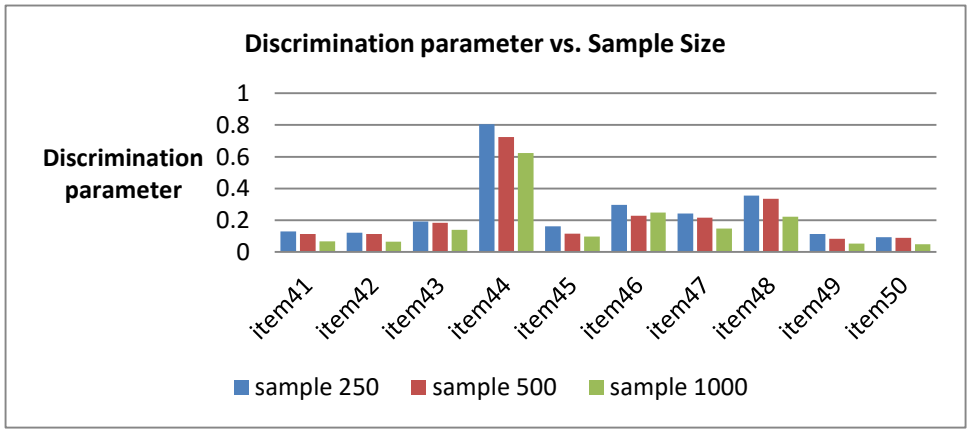


Figure 17: Item discrimination parameter estimates from 2plm

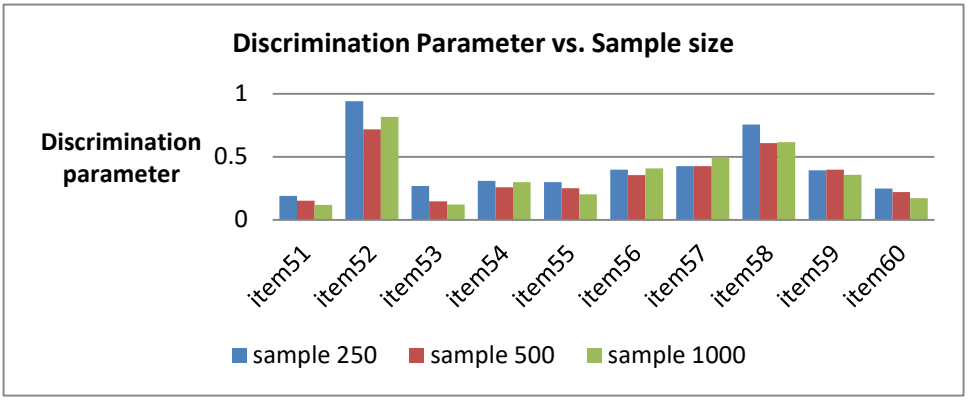


Figure 18: Item discrimination parameter estimates from 2plm

4.2.4 Graphical assessment of item difficulty estimates

Presented in this section is the graphical analysis conducted to inspect the behavior of difficulty parameter estimates from 2PL model for each item across the samples. In the plots it is visible that the difficulty parameter estimates are behaving in a similar manner across the samples. That is to say when the estimates for an item is increasing or decreasing, negative or positive it does so in all the three samples in most of the items. Therefore the researcher concluded that the item estimates across the samples were similar.

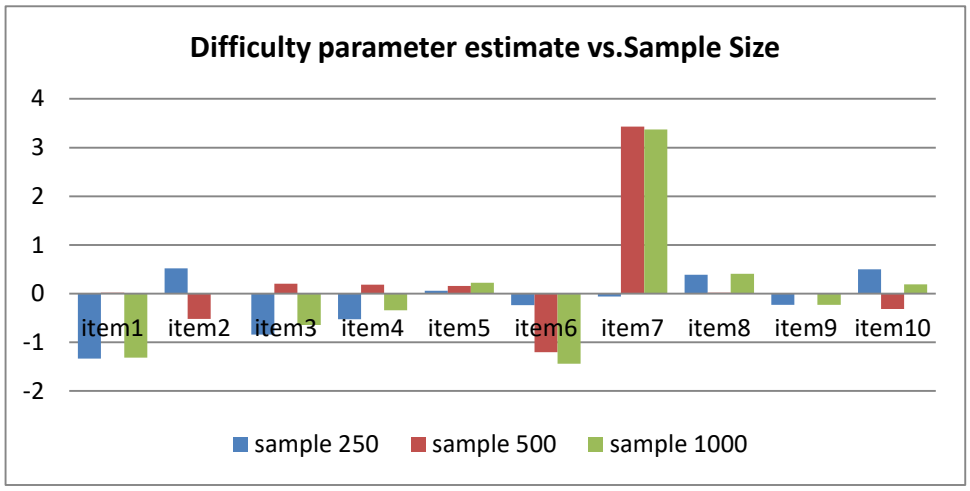


Figure 19: Item difficulty parameter estimates from 2plm

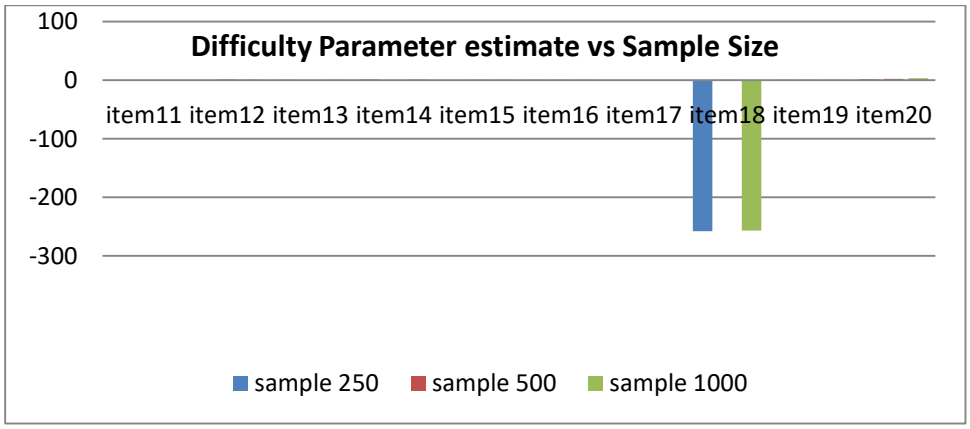


Figure 20: Item difficulty parameter estimates from 2plm

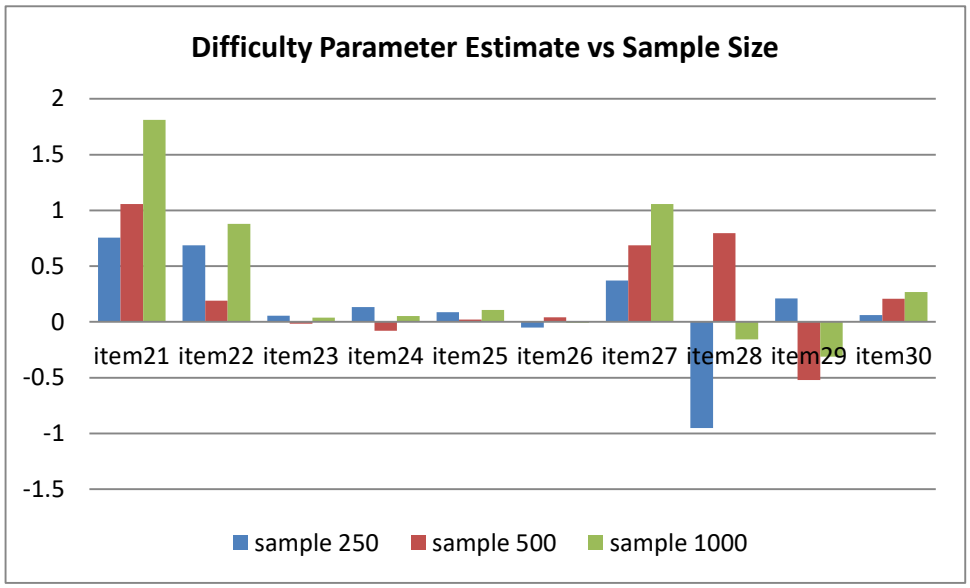


Figure 21: Item difficulty parameter estimates from 2plm

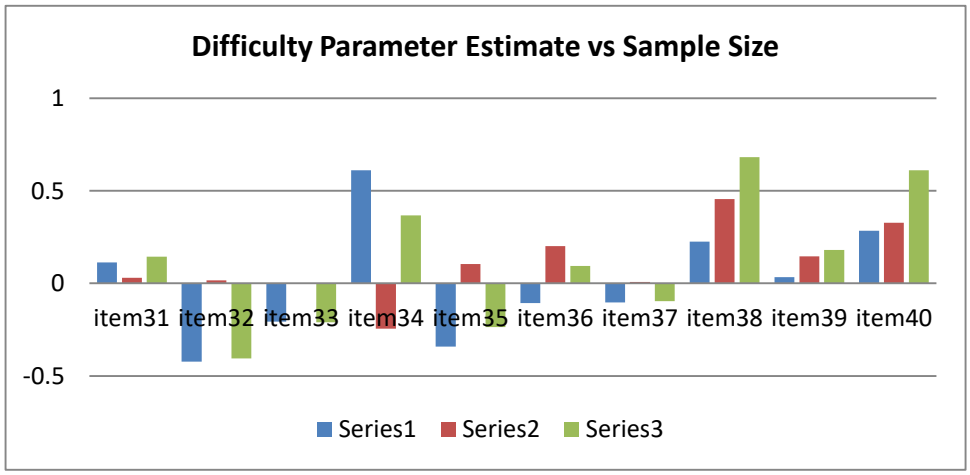


Figure 22: Item difficulty parameter estimates from 2plm

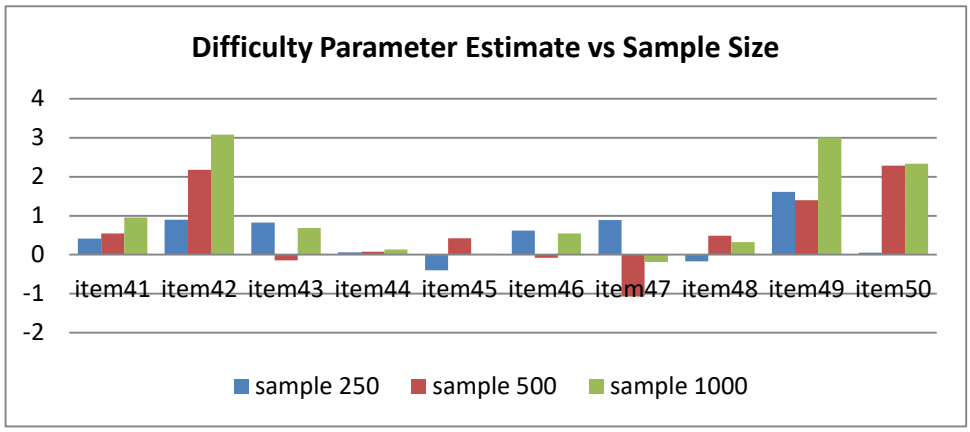


Figure 23: Item difficulty parameter estimates from 2plm

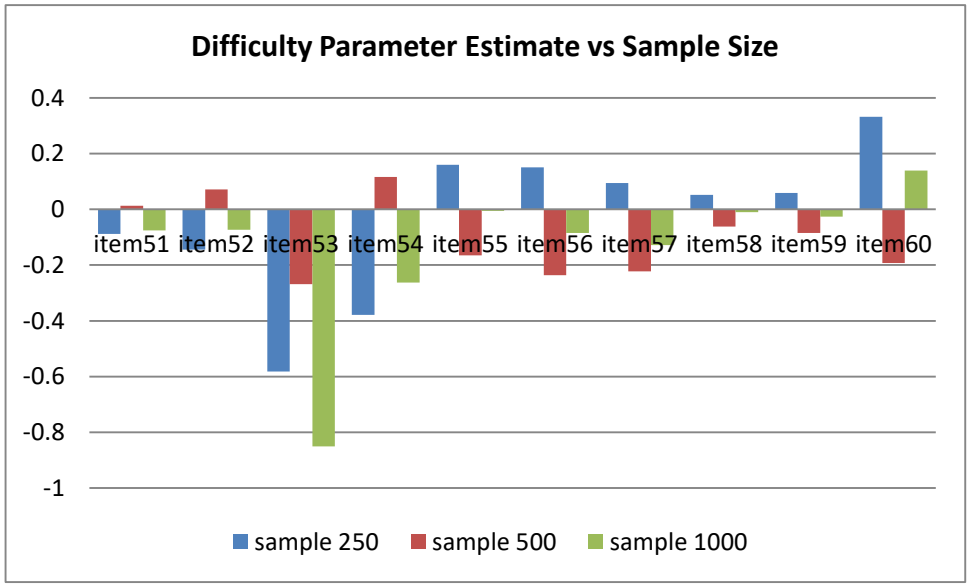


Figure 24: Item difficulty parameter estimates from 2plm

4.3 Comparison of item parameters in ANOVA

The section presents the F-test results of the one way ANOVA that were run on SPSS using the item and person parameters estimates, generated from examinee responses for the MANEB 2009 MSCE English language examination paper. Three independent

samples of (250, 500, and 1000) were drawn and parameters were generated in BILOG.MG3.The aim of the F-test was to find out whether the parameter estimates across three samples were statistically different or not.

4.3.1ANOVA results for 2plm item difficulty parameter

One hundred and eighty item difficulty parameter estimates were generated in 2PL model across the three sample sizes (250, 500 and 750) examinees. The parameters were then run on SPSS using F-test of one way ANOVA to compare the group means of the three sample sizes on item difficulty parameter estimates in order to determine if sample size has an effect on item difficulty parameter estimation. The tables 7 and 8 below present the descriptive statistics and the results of the F- test of one way ANOVA for the item

Difficulty parameter estimates.

Table 7:Descriptives for difficulty Parameter estimates

Samples	N	Mean	Std. Deviation	Std. Error
250	60	4.724117	33.6929972	4.3497472
500	60	.534383	1.7537535	.2264086
1000	60	.785700	2.1372121	.2759129
Total	180	2.014733	19.5038149	1.4537285

Table 7 provides familiar descriptive statistics “means and standard deviations” for the three independent sample size groups on of the dependent variable *(item difficulty parameter estimates)* for this analysis. The mean and standard deviation for the item difficulty parameter is higher for the small sample size of 250 examinees but this may be due to the effect of few outliers in the data. Therefore we cannot make solid conclusions based on descriptive statistics due to the pulling effect of this statistic.

Table 8: ANOVA Statistics for Item Difficulty Parameter estimate

Sources	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	662.563	2	331.282	.870	.421
Within Groups	67428.822	177	380.954		
Total	68091.385	179			

The main ANOVA summary table is divided into between group effects (effects due to the experiment) and within group effects (this is the unsystematic variation in the data).The between-group is the overall experimental effect. In this row we are told the sums of squares for the model (SSM =662.563). The sum of squares and mean squares represent the experimental effect. The row labeled *within group* gives details of the unsystematic variation within the data (the variation due to natural individual differences in the discrimination parameter estimates). The table tells us how much unsystematic variation exists (the residual sum of squares, SSR). It then gives the average amount of unsystematic variation, the mean squares (MSR). The test of whether the group means are the same is represented by the F-ratio for the combined between-group effect. The

value of this ratio is 0.870. Finally, SPSS tells us whether this value is likely to have happened by chance. The final column labeled *sig*. indicates how likely it is that an F-ratio of that size would have occurred by chance. In this case, there is a probability of 0. 412. An F-ratio of this size would have occurred by chance (that's only a 41.2% chance!). Social scientists use a cut of point of 0.05 (5%) as the criterion for statistical significance. Hence, because the observed significance value exceeds 0.05 we can say that the *three independent samples of* examinees (250, 500 and 1000) were not significantly different on *Item difficulty parameter estimates.* With, F (2,177) =0.870, sig=.412.

4.3.2 ANOVA results for item discrimination parameter estimates

One hundred and eighty discrimination parameter estimates were generated in 2pl model across the three sample sizes (250, 500 and 750) examinees. The parameters were then run on SPSS using F-test of one way ANOVA TO compare the group means of the three sample sizes on item discrimination parameter estimates in order to determine if sample size has an effect on discrimination parameter estimation. The tables 9 and 10 present the descriptive statistics and the results of the F- test of one way ANOVA for the discrimination parameter estimates.

Table 9: Descriptive Statistics for Item discrimination estimates

Samples	N	Mean	Std. Deviation	Std. Error
250	60	0.373650	0.2213559	0.0285769
500	60	0.347483	0.2217569	0.0286287
1000	60	0.332783	0.2475897	0.0319637
Total	180	0.351306	0.2298950	0.0171354

Table 9 provides familiar descriptive statistics "means and standard deviations" for the three independent sample size groups on of the dependent variable *(item discrimination parameter estimates)* for this analysis. The mean for the item discrimination parameter is slightly decreasing as the sample size increases from 250 examinees to 1000 examinees. However, we cannot make solid conclusions based on mean statistic due to the pulling effect of this statistic.

Table 10: ANOVA Statistics for Item discrimination estimates

Sources	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.051	2	0.026	0.484	0. 617
Within Groups	9.409	177	0.053		
Total	9.460	179			

The main ANOVA summary table is divided into between group effects (effects due to the experiment) and within group effects (this is the unsystematic variation in the data).The between-group is the overall experimental effect. In this row we are told the

sums of squares for the model (SSM =0.051). The sum of squares and mean squares represent the experimental effect. The row labeled *within group* gives details of the unsystematic variation within the data (the variation due to natural individual differences in the discrimination parameter estimates). The table tells us how much unsystematic variation exists (the residual sum of squares, SSR). It then gives the average amount of unsystematic variation; the mean squares (MSR). The test of whether the group means are the same is represented by the F-ratio for the combined between-group effect. The value of this ratio is 0.484. Finally, SPSS tells us whether this value is likely to have happened by chance. The final column labeled *sig*. indicates how likely it is that an F-ratio of that size would have occurred by chance. In this case, there is a probability of 0. 617. An F-ratio of this size would have occurred by chance (that's only a 6.17% chance!). Social scientists use a cut-off point of 0.05 (5%) as the criterion for statistical significance. Hence, because the observed significance value exceeds 0.05 we can say that the *three independent samples of* examinees (250, 500 and 1000) were not significantly different on *Item discrimination parameter estimates.* With, F (2,177) =0.484, sig=.617.

4.3.3 ANOVA results for person ability parameter estimates

One thousand seven hundred and fifty examinees' ability parameter estimates were generated in 2pl model across the three sample sizes (250,500 and 750) examinees. The parameters were then run on SPSS using F-test of one way ANOVA TO compare the group means of the three sample sizes on examinee's ability parameter estimates in order to determine if sample size has an effect on ability parameter estimation. The tables 11

and 12 present the descriptive statistics and the results of the F- test of one way ANOVA for the ability parameter estimates.

Table 11: Descriptive Statistics for examinee Parameter estimates

Samples	N	Mean	Std. Deviation
250	250	.000002	.9407372
500	500	.000000	.9353662
1000	1000	-.000001	.9360689
Total	1750	-.000001	.9360023

The summary table of descriptive statistics shows that the means and standard deviations for the person ability parameter estimates across the three samples are all most the same. However, the mean and standard deviations for the smallest sample of 250 examinees are slight higher than those of the larger samples of 500 and 1000 examinees respectively.

Table 12: ANOVA Statistics for examinee ability Parameter estimates

SOURCE	**SOME OF SQUARES**	**Df**	**Mean Square**	**F**	**Sig**
Between Groups	**.000**	**2**	**.000**	**.000**	**1.000**
Within Groups	**1533.175**	**1748**	**.877**		
Total	**1533.175**	**1750**			

The main ANOVA summary table is divided into between group effects (effects due to the experiment) and within group effects (this is the unsystematic variation in the data). The between-group is the overall experimental effect. In this row we are told the sums of squares for the model (SSM = 0.000). The sum of squares and mean squares represent the experimental effect. The row labeled *within group* gives details of the unsystematic variation within the data (the variation due to natural individual differences in the ability parameter estimates). The table tells us how much unsystematic variation exists (the residual sum of squares, SSR). It then gives the average amount of unsystematic variation; the mean squares (MSR). The test of whether the group means are the same is represented by the F-ratio for the combined between-group effect. The value of this ratio is 0.000. Finally, SPSS tells us whether this value is likely to have happened by chance. The final column labeled *sig*. Indicates how likely it is that an F-ratio of that size would have occurred by chance. In this case, there is a probability of 1.000. An F-ratio of this size would have occurred by chance (that's a 100% chance!). Social scientists use a cut of point of 0.05 (5%) as the criterion for statistical significance. Hence, because the observed significance value exceeds 0.05 we can say that the *three independent samples of* examinees (250, 500 and 1000) were not significantly different on *examinees 'ability parameter estimates.* With, F (2,177) =0.000, sig=.1.000.

4.4 Summary for the results of the study

In summary the model data fit in the 1PL model was poor as compared to 2PL model. The data fit was relatively good though not to the same extent as the fit from most of

simulation studies. The study therefore proceeded to examine parameters generated using 2PLM across the independent samples of sizes: 250, 500, and 1000. The results from comparing the corresponding item and examinees 'parameters estimates within the model showed that the parameters were statistically equivalent across the three samples.

CHAPTER FIVE

DISCUSSION, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

5.0 Chapter over view

This final section of the thesis will briefly review and summarise the main results found in chapter 4. This chapter describes the results of the study in four sections. The first section discusses the findings of the study from the research questions. The next section compares the results of this study to findings from previous research. The third section discusses implications of the findings for practice and policy and discusses the limitations of the study. Final part presents recommendations and conclusions.

5.1 Discussion

The first question that the study sought to address was "Which IRT model fits the data" this question intended to help the researcher in selecting the appropriate model for generating the item and person parameter estimates. Through this preliminary analysis the 2PL model was chosen because it fitted the data well than the other models.

The second research question compared the item difficulty estimates from the three samples of varied sizes of: 250, 500 and 1000. The results showed that the item difficulty parameter estimates were not statistically different across the three samples. This led to

accepting the null hypothesis that “Differences in Groups with varied sample sizes have no statistically significant influence on the item difficulty parameter estimates generated using 2PL model. Theoretically this finding supports the principle of item invariance which is the cornerstone of the IRT framework which says that item parameters across different samples of examinees must be equivalent.

The third research question compared the item discrimination parameter estimates from the three samples with varied sizes of: 250, 500 and 1000. The results showed that the item discrimination parameter estimates were not statistically different across these samples and this led us to accepting the null hypothesis that “Differences in sample sizes have no statistically significant influence on the item difficulty parameter estimates “**a**” based 2PL IRT model. This affirms the theory that, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically by (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and has been widely accepted within the measurement community.

The last research question examined the differences in person ability parameter estimates from 2P model based on three samples of varied sizes of: 250, 500 and 1000.the results showed that the person ability parameter estimates were not statistically different across these sample sizes therefore this lead us to accepting the null hypothesis that “Differences in sample sizes have no statistically significant influence on the person ability parameter

estimates (θ) based on 2Pl IRT model. This affirms the theory that, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and has been widely accepted within the measurement community.

5.2 Relationship of the findings to prior research

The past research studies though most of them used simulated data and employed Root Mean Squared Errors (RMSEs) as evaluative criteria for assessing the influence of small samples on item parameter estimates are similar to the findings of this study.

In an analysis of the effect of sample size on linear equating, Ree and Jensen (1983) examined several combinations of calibration and equating sample sizes. They suggested a minimum sample size of 500.Hambleton and Cook (1983) simulated tests of 10, 20, and 80 items with sample sizes of 50, 200, and 1000 in order to determine the effect of sample size on the standard errors of ability estimation curves. Ability scores were drawn from a standard normal distribution, and item parameters were estimated using heuristic estimation software (Urry, 1974). They concluded that adequate precision could be obtained near the center of the ability continuum under most testing conditions with a sample size of 200 which is comparable to the 250 sample size that was used in this study.

Lim and Drasgow (1990) examined the parameter recovery capabilities of BILOG for samples of 250 examinees on a 20-item test. They reported that Bayes modal estimates showed less estimation error when sample size (n) = 250. Michael R. Harwell and Janine E. Janosky (1991) Effects of Small Datasets and Varying Prior Variances on Item Parameter and found that once samples size exceeds 250, the estimation error tends to be reasonable.

These findings from previous studies have been supported by the findings of this present study which is also pointing at a sample of size 250 examinees as being reasonable sample size to be used in generating stable item and person parameter estimates.

5.3 Implications for practice and policy

When we give a test, it is usually because we have to make a decision and we want the results of the testing situation to help us make that decision. We have to interpret those results, and to make the case that our interpretations are valid for that situation. Validity, therefore, is an argument that we make about our assumptions, based on test scores. We must make the case that the instrument we use does, in fact, measure the psychological trait we hope to measure. Validity is, according to the Standards for Educational and Psychological Testing, "the most fundamental consideration in developing and evaluating tests" (cited in Hogan & Agnello, 2004).

One kind of support for the validity of the interpretation is that the test measures the psychological trait consistently. This is known as the reliability of the test. Reliability, i.e., a measure of the consistency of the application of an instrument to a particular

population at a particular time, is a necessary condition for validity. A reliable test may or may not be valid, but an unreliable test can never be valid. This means that a test cannot be more valid than it is reliable, i.e., reliability is the upper limit of validity. It is important to remember that any instrument, i.e., the MANEB, SLEP test or TOEFL, does not have "reliability." An instrument that demonstrates high reliability in one situation may show low reliability in another. Reliability resides in the interaction between a particular task and a particular population of test-takers. This study has examined the item parameter invariance through the interaction of instrument with the sample size. The findings of the study showed that item parameters from the sample of size 250 are statistically equivalent as those produced from samples of sizes 500 or 1000.

These results will help to inform policy, in future, examination boards and other stake holders may reduce pretesting cost for smaller sample of 250 examinees will be sufficient. These findings will also improve test security by reducing item exposure since fewer examinees need to see each item to estimate the item parameters accurately.

In practice this study contributes in determining and establishing the minimum sample size which can be employed when generating item parameter estimates in IRT 1P and 2P models.

In the academic circles the findings of this research has increase the empirical knowledge on the influence of sample size on item parameter estimates based on IRT 1P and 2P theoretical framework.

5.4 Limitations of the current study

The first shortcoming of the investigation is the limited item pool used in the study. Although the examinee pool is quite adequate in the sense that a variety of different samples can be drawn from it, the same cannot be said about the item pool. Ideally, the test item pool should be larger and more diverse in terms of item characteristics (including both homogenous and heterogeneous items) so that items can be sampled from the pool to study the behaviors of IRT item statistics under different conditions of item characteristics. Future studies may benefit from using several different testing databases.

In this study, unlike in most of simulation situations where most of the items, by design, fit the 3PL model well, the real data was messier, had a very poor fit to the 3PL model, hence the researcher could not proceed to work with 3P model. Additionally, this study may not generalise to other IRT models

5.5 Recommendations

This study employed ANOVAs to assess the equivalence of item parameter estimates across varied sample sizes, it may be important to examine these samples using other evaluative criteria like amount of Item differential item function, item Bias and Root MEAN Squared Errors.

While this study examined the effect of sample sizes on statistical equivalence item parameter estimates, it is also important to understand its effect on ability parameter estimates and standard and measurement errors.

References

Abdel-Fattah, A.-f. A. (April, 1994). *Comparing BILOG and LOGIST estimates for normal, truncated normal and beta ability distributions.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Allison, P. D. (2001). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage Publications, Inc.

Assessment Systems Corporation (1987). MICROCAT: A computer program for computerised adaptive testing (2nd ed.) (Computer Software): Assessment Systems Corporation.

Assessment Systems Corporation (1989). MICROCAT: A computer program for computerised adaptive testing (3rd ed.).(Computer Software): Assessment Systems Corporation.

Baker, F., & Kim, S.-H. (2004). *Item response theory parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*, 111-141.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Barnes, L. L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*, 143-157.

Bebbington, A.C. 1975. A simple method of drawing a sample without replacement. *Applied Statistics*, 24(1).

Berger, M., King, C.Y.J., & Wong, W. (2000). Minimax d-optimal designs for item response theory models. *Psychometrika, 65*, 377-390. 122.

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

Box, G. E. P. (1979), "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N., *Robustness in Statistics*, Academic Press, pp. 201–236

Chris Wheadon et al (2013).The effect of sample size on item parameter estimation for the partial credit model Journal: *Int. J. of Quantitative Research in Education*, 1(3), 297-315.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Croll, P. R., & Urry, V. W. (1978). ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options - Version 78.5 (Computer Software). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

De Mars, C. (2001). Group differences based on IRT scores: Does the model matter? *Educational and Psychological Measurement, 61*, 60-70.

Dodeem H (2004). The relationship between item parameters and item fit. J. Educa.Meas. 41(3):261-270.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.123

Foley, B. F. (2009, April). *Improving IRT item parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique*. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Gifford, J., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*, 33-43.

Gierl, M. J., & Ackerman, T. (1996). Software review: XCALIBRETM marginal maximum likelihood estimation program, Windows version 1.10. *Applied Psychological Measurement, 20*, 303-307.

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

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

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. Linn (Ed.). *Educational measurement* (3rd ed.), (pp. 147-200). Washington, D.C.: American Council on Education.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31–49). New York, NY: Academic Press.

Harvey, R., & Hammer, A. (1999). Item Response Theory. 27 (3), 353-383.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Hwang, D.-Y. (2002). *Classical test theory and item response theory: Analytical and empirical comparisons.* Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Jensema, C. (1972). *An application of latent trait mental test theory to the Washington Pre College Testing Battery (Research Bulletin)*. Seattle, WA: University of Washington, Bureau of Testing.

Jensema, C. (1976). *A simple technique for estimating latent trait mental test parameters.*

Educational and Psychological Measurement, 36, 705-715. 125.

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. In S. M. Downing & T. M. Haladyna (Eds.), *Hand bookof test development* (pp. 487-525). Mahwah, N.J.: L. Erlbaum.

Kim, S.-H. (2007). Some posterior standard deviations in item response theory. *Educational and Psychological Measurement, 67*, 258-279.

Kirisci, L., Hsu, T.-c., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146 162.

Liang,T., Han, K. T., & Hambleton, R.K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement, 33*(5), 411-412.

Little, R.J.A., Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd ed.). New

York, NY: Wiley.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F. (1980). *Applications of item response theory to practical testing problems*.

Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48*, 425-435.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921-943. 126

Mislevy, R. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*,177 195.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items [Computer software]. Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1986). PC-BILOG: Item analysis and test scoring with binary logistic models (Computer software). Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1989). PC-BILOG 3: Item analysis and test scoring with binary logistic models (Computer software). Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1997). BILOG 3: Item analysis and test scoring with binary logistic models (Computer software). Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating-scale data (Computer software). Chicago: Scientific Software.

Parshall, C. G., Kromrey, J. D., Chason, W. M., & Yi, Q. (1997). *Evaluation of parameter estimation under modified IRT models and small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN Parshall, C. G.,

Kromrey, J. D., & Chason, W. M. (1996). *Comparison of alternative models for item parameter estimation with small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.127.

Patsula, L. N., & Gessaroli, M. E. (1995). *A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Parshall, C. G., Kromrey, J. D., & Chason, W. M. (1996). *Comparison of alternative models for item parameter estimation with small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Ramsay, J. O. (1993). TESTGRAF: *A program for the graphical analysis of multiple choice test data (Computer software)*. Montreal: McGill University.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3,* 371-385.

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss & R. D. Bock (Eds.), *New horizons in testing: Latent trait test theory and computerised adaptive testing* (pp. 135-146). New York: Academic Press.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rupp A. A., Zumbo B. D. (2006). Understanding parameter invariance in unidimensional IRT models. Educ. Psychol. Meas. 66, 63–8410.1177/0013164404273942

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: J. Wiley & Sons.

Scherbaum, Cohen-Charash, & Kern (2006). Educational and Psychological Measurement, 66, 1047-1063.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*, 299-311.

Setiadi, H. (1997). *Small sample IRT item parameter estimates.* Unpublished Ed.D., University of Massachusetts Amherst, United States, Massachusetts.

Sireci, S. G. (1992). *The utility of IRT in small-sample testing applications.* Paper presented at the Annual Meeting of the American Psychological Association, Washington, DC.

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-Bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13,* 391-402.

Stone, C. A., Weissman, A., & Lane, S. (2005). The consistency of student proficiency classifications under competing IRT models. *Educational Assessment, 10,* 125-146.

Stocking, M. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika, 55,* 461-475.

Streubert, H.J., & Carpenter, D.R.(Eds.). (1999). Qualitative research in nursing. Advancing the humanistic imperative (2nd ed.). Philadelphia: Lippincott.

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss & R. D. Bock (Eds.), *New horizons in testing: latent trait test theory and computerized adaptive testing* (pp. 13-30). New York: Academic Press.

Swaminathan, H., & Gifford, J. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27,* 27-51.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (2nd ed.), (pp. 159-184). Westport, CT: Praeger Publishers.

Timminga, E. (1995). Optimum examinee samples for item parameter estimation in item response theory: A multi-objective programming approach. *Psychometrika, 60,* 137-154.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses.*Educational and Psychological Measurement, 34,* 253-269.

Urry, V. W. (1977). OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options (Computer software). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center.

Urry, V. W. (1978). ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options (Computer software). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center.

Vale, C. D., Maurelli, V.A., Gialluca, K. A., Weiss, D.J., & Ree, M.J. (1981). *Methods for linking item parameters* (AFHRL-TR-81-10). Brooks Air Force Base TX: U.S. Air Force Human Resources Laboratory, Manpower and Personnel Division.

Vale, C. D., & Gialluca, K. A. (1985). ASCAL: A microcomputer program for estimating logistic IRT item parameters (ONR-85-4) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement, 12,* 53-67.

Wainer, H. & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), *Computer adaptive testing: A primer (*2nd ed.), (pp. 271-299). Westport, CT: Praeger Publishers.

Wendler, C.L & Walker, M.E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445-467). Mahwah, N.J.: L. Erlbaum.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide (Computer software). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1985). LOGIST user's guide (LOGIST 5 version 2.1) (Computer software). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347-364.

Wood, R., Wingersky, M. S., & Lord, F. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6) (Computer Software). Princeton, NJ: Educational Testing Service.

Wright, B., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yen, W. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*(2), 275-291.

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), pp. 111-153). Westport, CT: Praeger Publishers.

Yoes, M. E. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the three-parameter logistic item response theory model. (Volumes I and II).* Unpublished doctoral thesis, University of Minnesota, Minneapolis/St. Paul, MN.

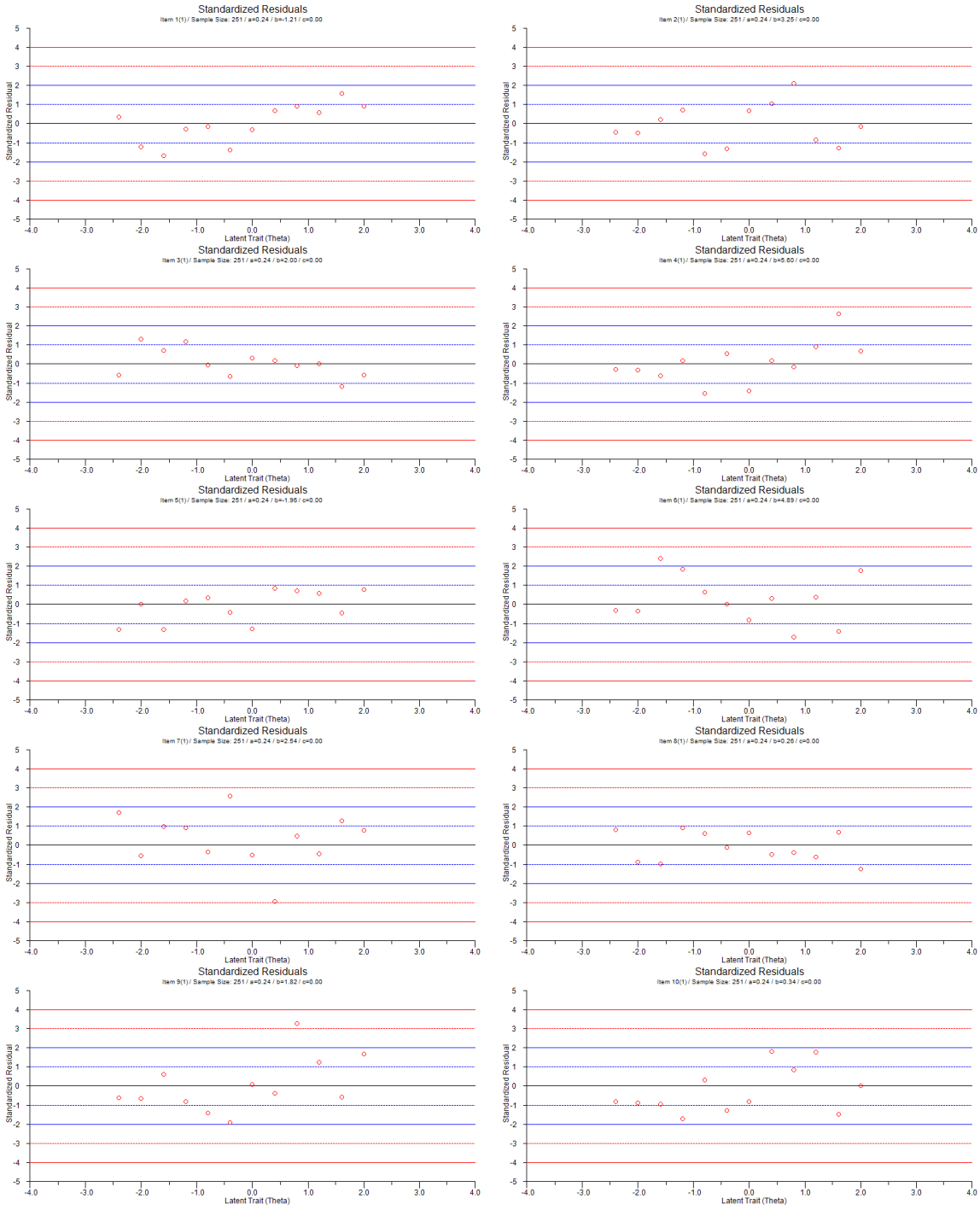
Yoes, M. E. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model (ASC Technical Rep. 95- 1R).* Saint Paul, MN: Assessment Systems Corporation.

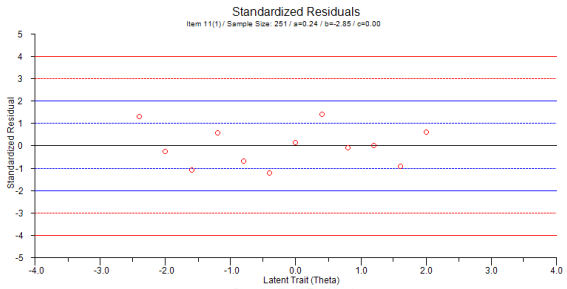
Yoes, M. E. (1996). User's manual for the XCALIBRE marginal maximum-likelihood estimation program (Computer software). St. Paul, MN: Assessment Systems Corp.

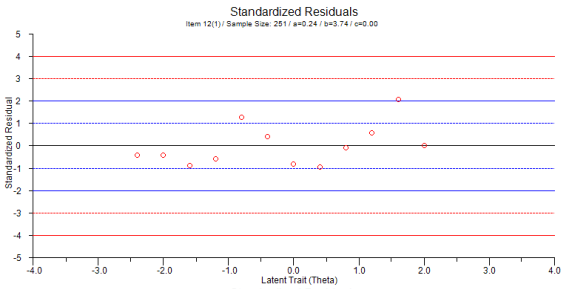
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (2003*). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* (Computer software). Lincolnwood, IL: Scientific Software International, Inc

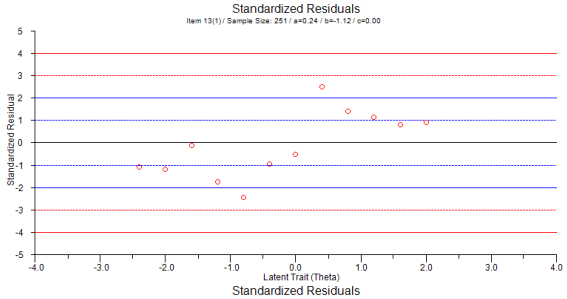
List of Appendices

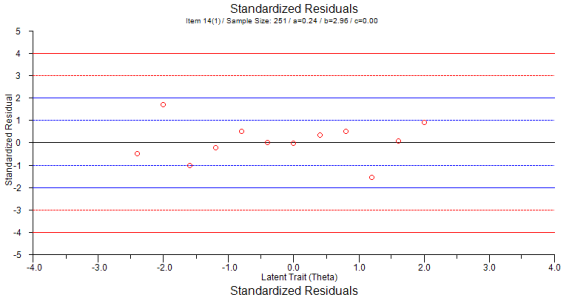
APPENDIX A: SRs from 1plm for sample 250

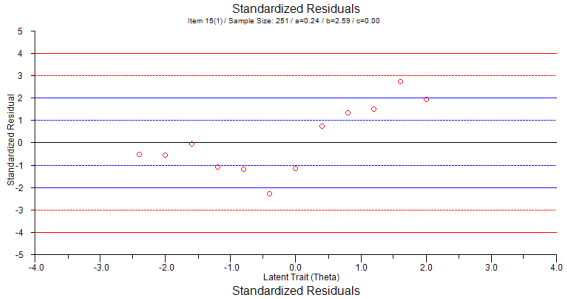


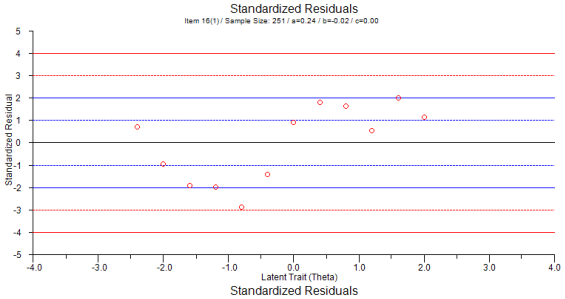

Standardized Residuals
Item 11(1) / Sample Size: 251 / a=0.24 / b=-2.65 / c=0.00
Standardized Residual
Latent Trait (Theta)

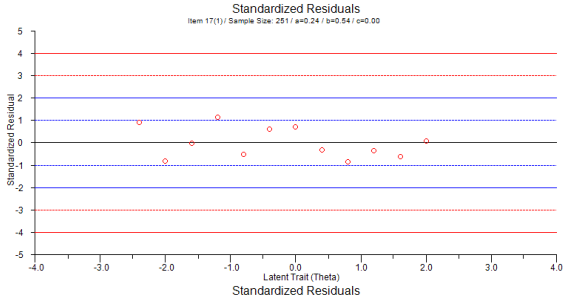

Standardized Residuals
Item 12(1) / Sample Size: 251 / a=0.24 / b=3.74 / c=0.00
Standardized Residual
Latent Trait (Theta)

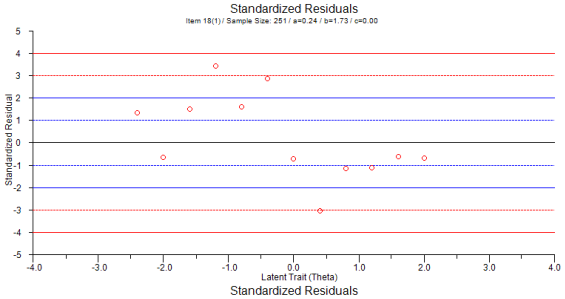

Standardized Residuals
Item 13(1) / Sample Size: 251 / a=0.24 / b=-1.12 / c=0.00
Standardized Residual
Latent Trait (Theta)

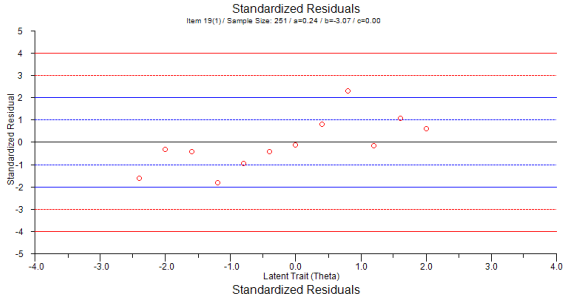

Standardized Residuals
Item 14(1) / Sample Size: 251 / a=0.24 / b=2.96 / c=0.00
Standardized Residual
Latent Trait (Theta)

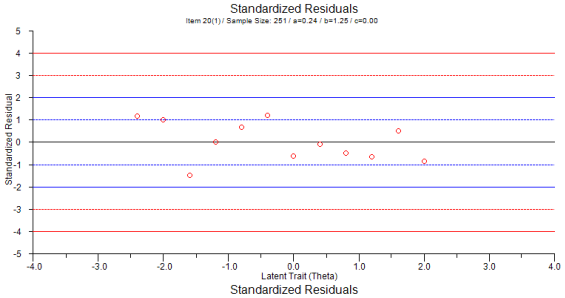

Standardized Residuals
Item 15(1) / Sample Size: 251 / a=0.24 / b=2.59 / c=0.00
Standardized Residual
Latent Trait (Theta)

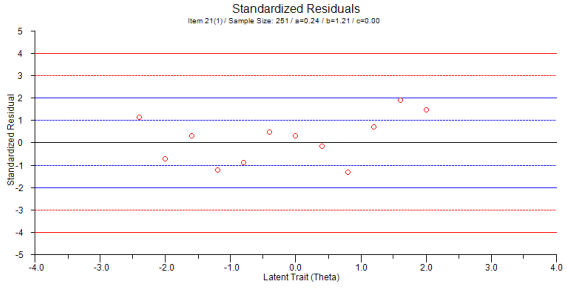

Standardized Residuals
Item 16(1) / Sample Size: 251 / a=0.24 / b=-0.02 / c=0.00
Standardized Residual
Latent Trait (Theta)

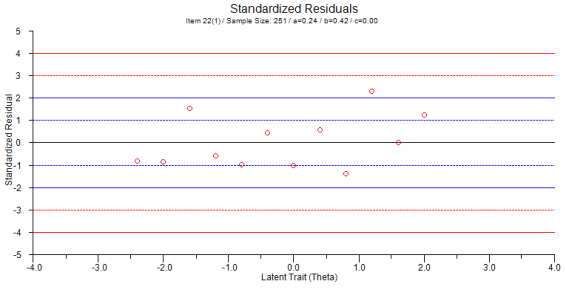

Standardized Residuals
Item 17(1) / Sample Size: 251 / a=0.24 / b=0.54 / c=0.00
Standardized Residual
Latent Trait (Theta)

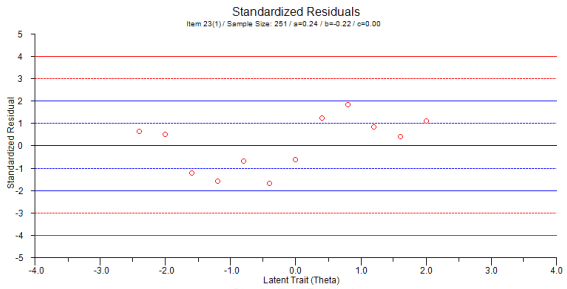

Standardized Residuals
Item 18(1) / Sample Size: 251 / a=0.24 / b=1.73 / c=0.00
Standardized Residual
Latent Trait (Theta)

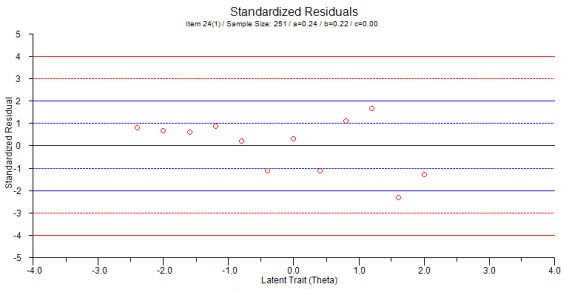

Standardized Residuals
Item 19(1) / Sample Size: 251 / a=0.24 / b=-3.07 / c=0.00
Standardized Residual
Latent Trait (Theta)

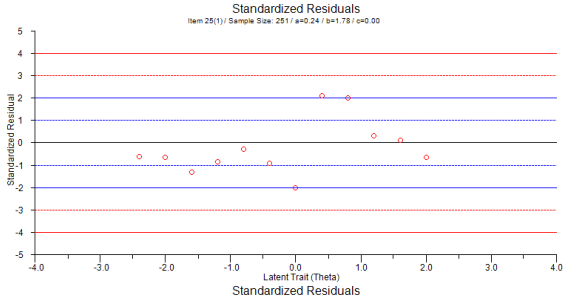

Standardized Residuals
Item 20(1) / Sample Size: 251 / a=0.24 / b=1.25 / c=0.00
Standardized Residual
Latent Trait (Theta)

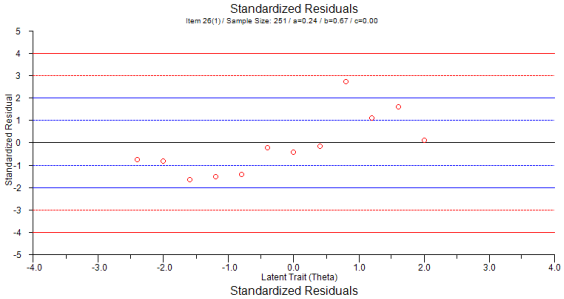

Standardized Residuals
Item 21(1) / Sample Size: 251 / a=0.24 / b=1.21 / c=0.00
Standardized Residual
Latent Trait (Theta)

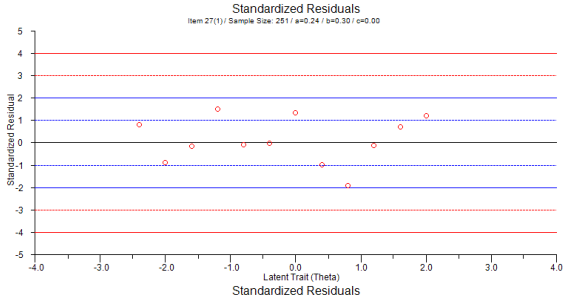

Standardized Residuals
Item 22(1) / Sample Size: 251 / a=0.24 / b=0.42 / c=0.00
Standardized Residual
Latent Trait (Theta)

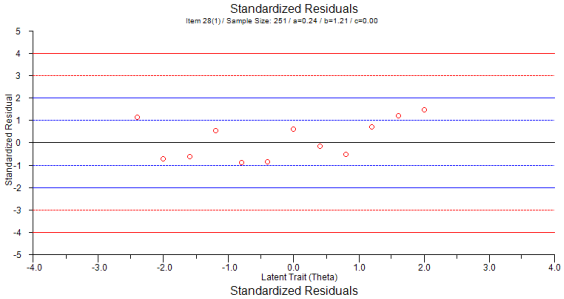

Standardized Residuals
Item 23(1) / Sample Size: 251 / a=0.24 / b=-0.22 / c=0.00
Standardized Residual
Latent Trait (Theta)

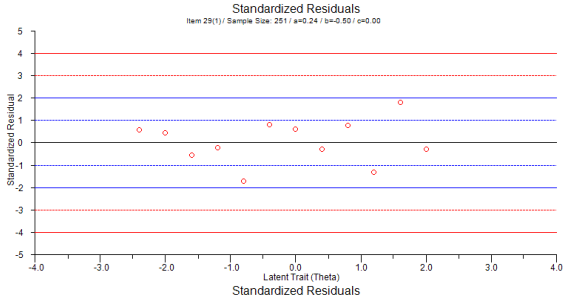

Standardized Residuals
Item 24(1) / Sample Size: 251 / a=0.24 / b=0.22 / c=0.00
Standardized Residual
Latent Trait (Theta)

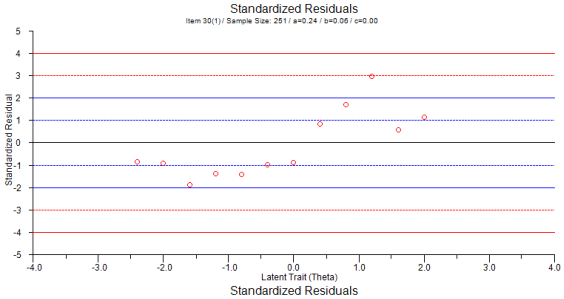

Standardized Residuals
Item 25(1) / Sample Size: 251 / a=0.24 / b=1.78 / c=0.00
Standardized Residual
Latent Trait (Theta)

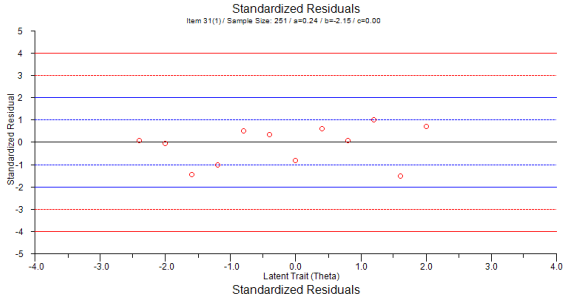

Standardized Residuals
Item 26(1) / Sample Size: 251 / a=0.24 / b=0.67 / c=0.00
Standardized Residual
Latent Trait (Theta)

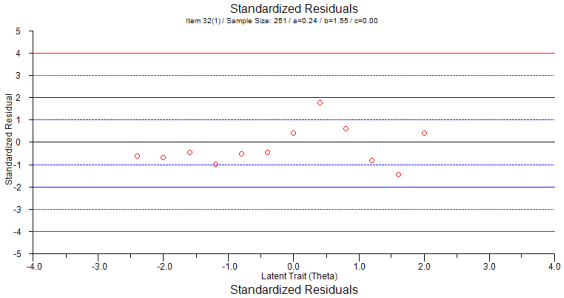

Standardized Residuals
Item 27(1) / Sample Size: 251 / a=0.24 / b=0.30 / c=0.00
Standardized Residual
Latent Trait (Theta)

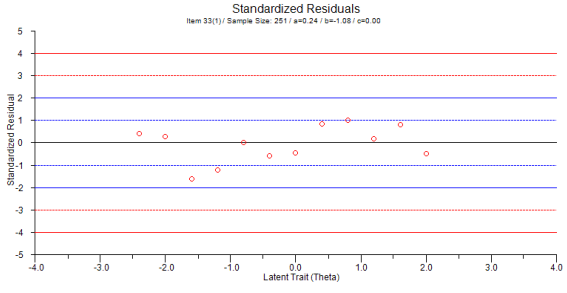

Standardized Residuals
Item 28(1) / Sample Size: 251 / a=0.24 / b=1.21 / c=0.00
Standardized Residual
Latent Trait (Theta)

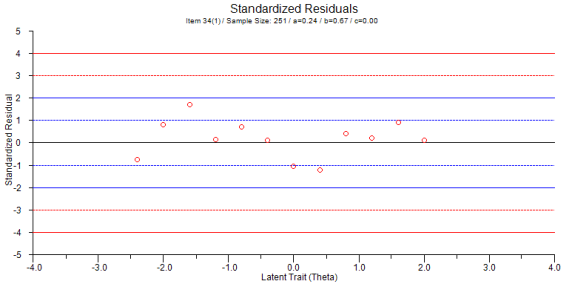

Standardized Residuals
Item 29(1) / Sample Size: 251 / a=0.24 / b=-0.50 / c=0.00
Standardized Residual
Latent Trait (Theta)

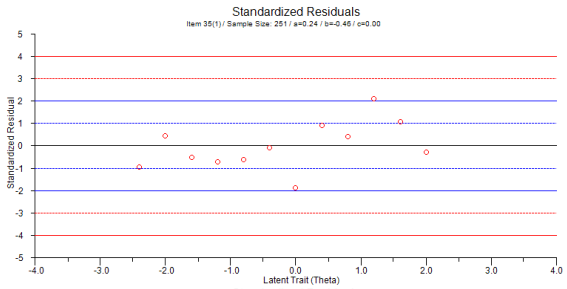

Standardized Residuals
Item 30(1) / Sample Size: 251 / a=0.24 / b=0.06 / c=0.00
Standardized Residual
Latent Trait (Theta)

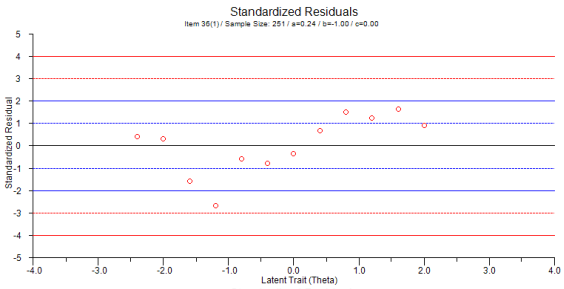

Standardized Residuals
Item 31(1) / Sample Size: 251 / a=0.24 / b=-2.15 / c=0.00
Standardized Residual
Latent Trait (Theta)

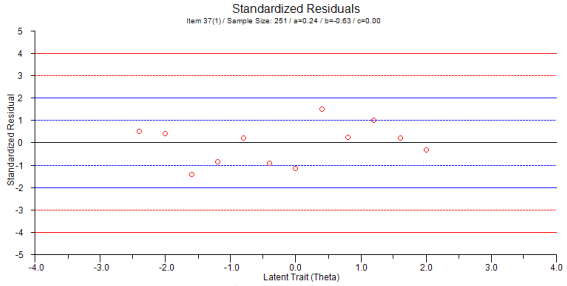

Standardized Residuals
Item 32(1) / Sample Size: 251 / a=0.24 / b=1.55 / c=0.00
Standardized Residual
Latent Trait (Theta)

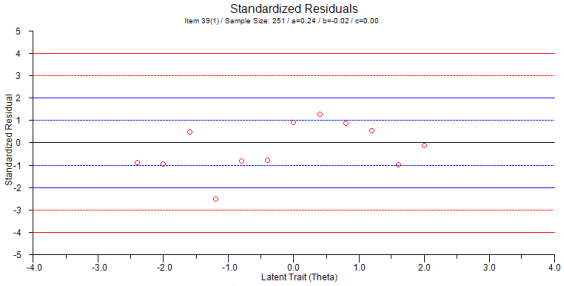

Standardized Residuals
Item 33(1) / Sample Size: 251 / a=0.24 / b=-1.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

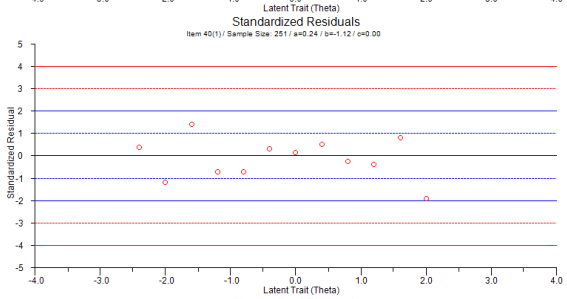

Standardized Residuals
Item 34(1) / Sample Size: 251 / a=0.24 / b=0.67 / c=0.00
Standardized Residual
Latent Trait (Theta)

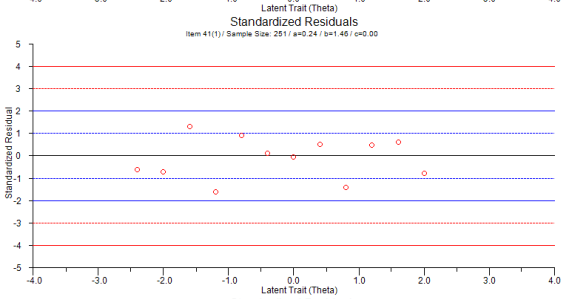

Standardized Residuals
Item 35(1) / Sample Size: 251 / a=0.24 / b=-0.45 / c=0.00
Standardized Residual
Latent Trait (Theta)

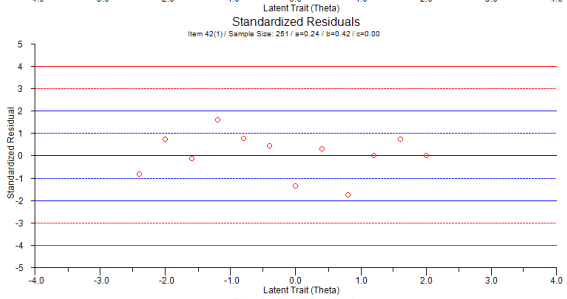

Standardized Residuals
Item 36(1) / Sample Size: 251 / a=0.24 / b=-1.00 / c=0.00
Standardized Residual
Latent Trait (Theta)

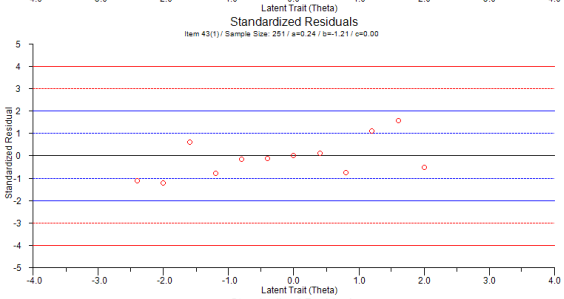

Standardized Residuals
Item 37(1) / Sample Size: 251 / a=0.24 / b=-0.63 / c=0.00
Standardized Residual
Latent Trait (Theta)

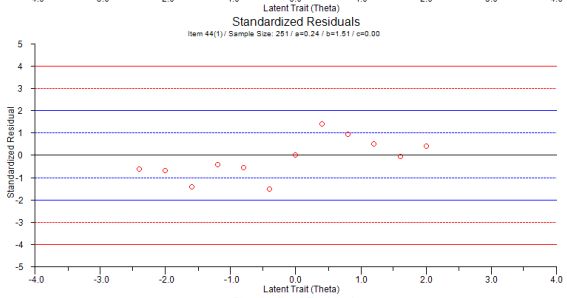

Standardized Residuals
Item 39(1) / Sample Size: 251 / a=0.24 / b=-0.02 / c=0.00
Standardized Residual
Latent Trait (Theta)

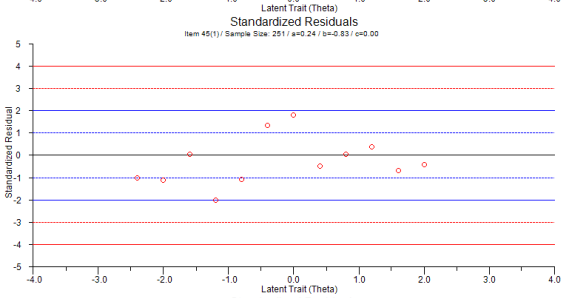

Standardized Residuals
Item 40(1) / Sample Size: 251 / a=0.24 / b=-1.12 / c=0.00
Standardized Residual
Latent Trait (Theta)

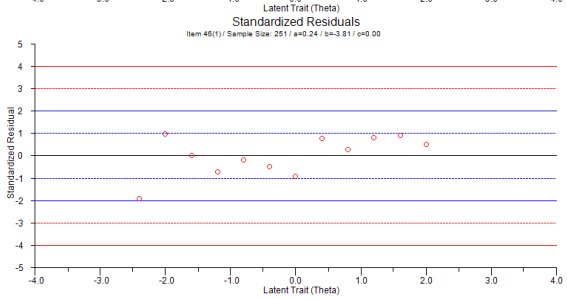

Standardized Residuals
Item 41(1) / Sample Size: 251 / a=0.24 / b=1.46 / c=0.00
Standardized Residual
Latent Trait (Theta)

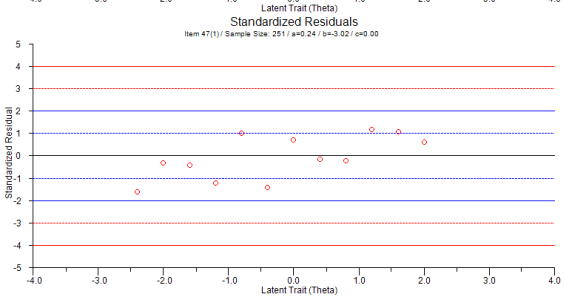

Standardized Residuals
Item 42(1) / Sample Size: 251 / a=0.24 / b=0.42 / c=0.00
Standardized Residual
Latent Trait (Theta)

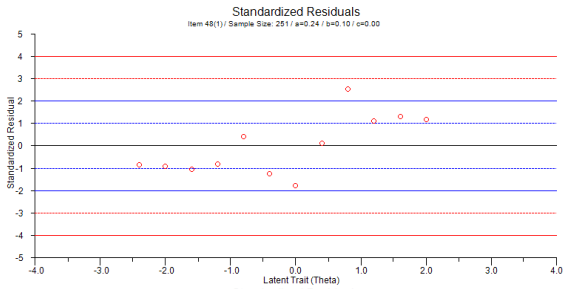

Standardized Residuals
Item 43(1) / Sample Size: 251 / a=0.24 / b=-1.21 / c=0.00
Standardized Residual
Latent Trait (Theta)

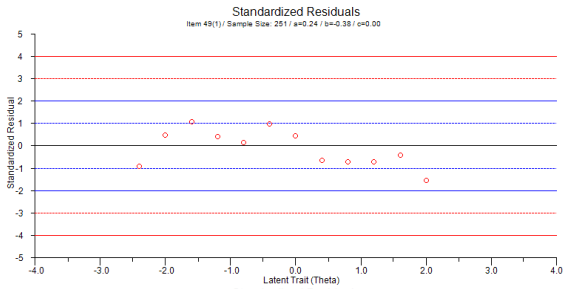

Standardized Residuals
Item 44(1) / Sample Size: 251 / a=0.24 / b=1.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

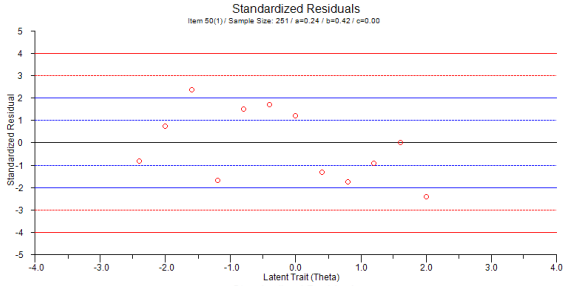

Standardized Residuals
Item 45(1) / Sample Size: 251 / a=0.24 / b=-0.83 / c=0.00
Standardized Residual
Latent Trait (Theta)

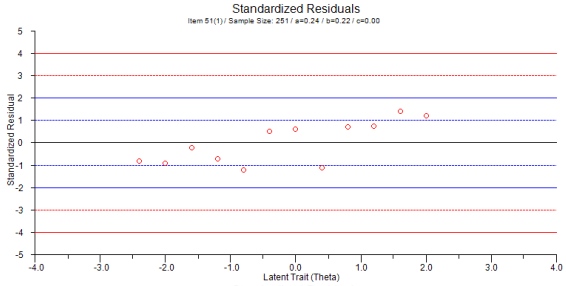

Standardized Residuals
Item 46(1) / Sample Size: 251 / a=0.24 / b=-3.81 / c=0.00
Standardized Residual
Latent Trait (Theta)

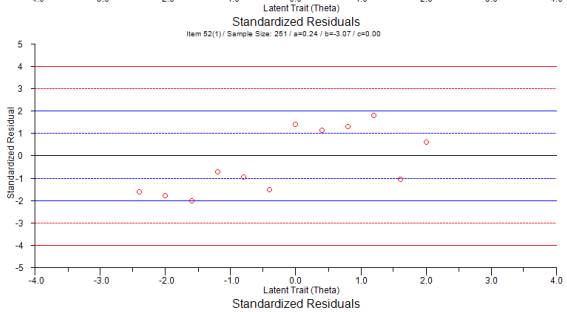

Standardized Residuals
Item 47(1) / Sample Size: 251 / a=0.24 / b=-3.02 / c=0.00
Standardized Residual
Latent Trait (Theta)

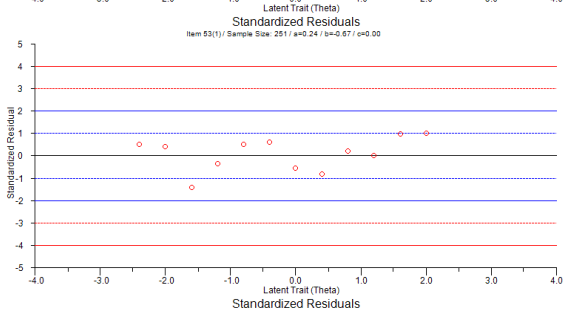

Standardized Residuals
Item 48(1) / Sample Size: 251 / a=0.24 / b=0.10 / c=0.00
Standardized Residual
Latent Trait (Theta)

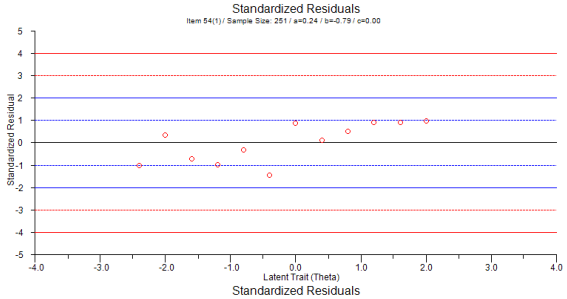

Standardized Residuals
Item 49(1) / Sample Size: 251 / a=0.24 / b=-0.38 / c=0.00
Standardized Residual
Latent Trait (Theta)

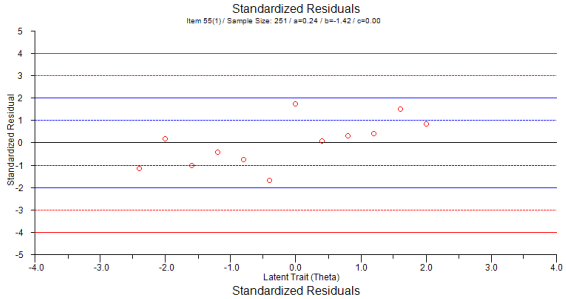

Standardized Residuals
Item 50(1) / Sample Size: 251 / a=0.24 / b=0.42 / c=0.00
Standardized Residual
Latent Trait (Theta)

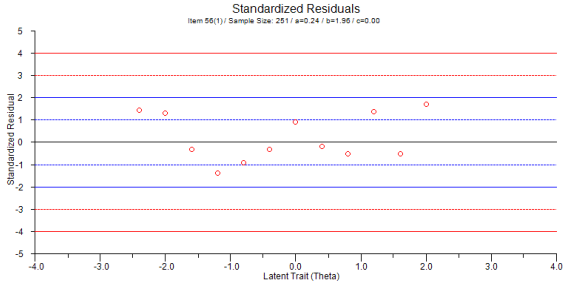

Standardized Residuals
Item 51(1) / Sample Size: 251 / a=0.24 / b=0.22 / c=0.00
Standardized Residual
Latent Trait (Theta)

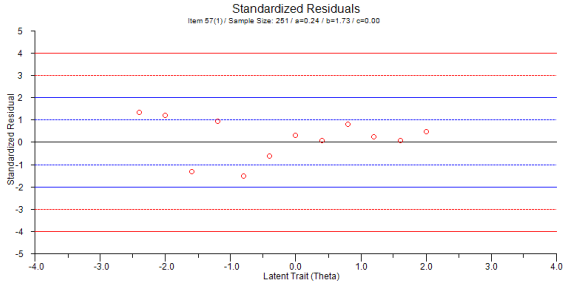

Standardized Residuals
Item 52(1) / Sample Size: 251 / a=0.24 / b=-3.07 / c=0.00
Standardized Residual
Latent Trait (Theta)

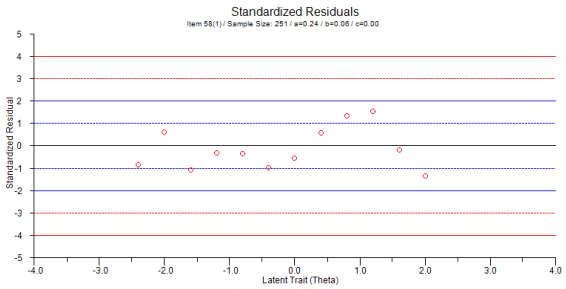

Standardized Residuals
Item 53(1) / Sample Size: 251 / a=0.24 / b=-0.67 / c=0.00
Standardized Residual
Latent Trait (Theta)

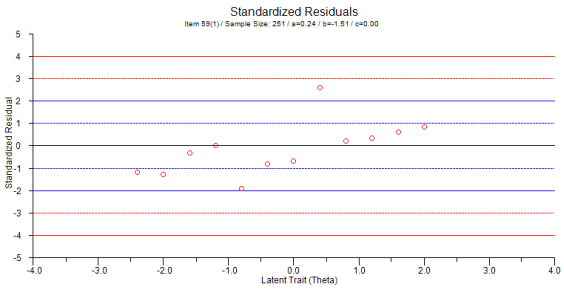

Standardized Residuals
Item 54(1) / Sample Size: 251 / a=0.24 / b=-0.79 / c=0.00
Standardized Residual
Latent Trait (Theta)

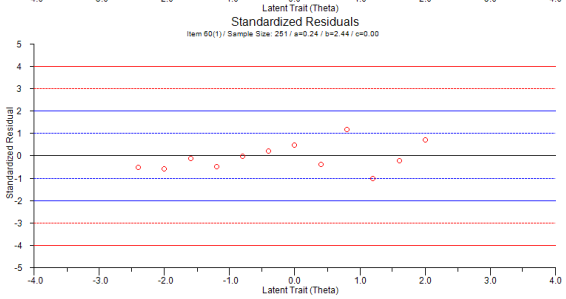

Standardized Residuals
Item 55(1) / Sample Size: 251 / a=0.24 / b=-1.42 / c=0.00
Standardized Residual
Latent Trait (Theta)


Standardized Residuals
Item 56(1) / Sample Size: 251 / a=0.24 / b=1.96 / c=0.00
Standardized Residual
Latent Trait (Theta)

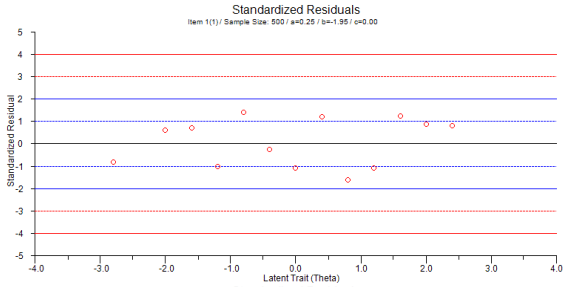

Standardized Residuals
Item 57(1) / Sample Size: 251 / a=0.24 / b=1.73 / c=0.00
Standardized Residual
Latent Trait (Theta)

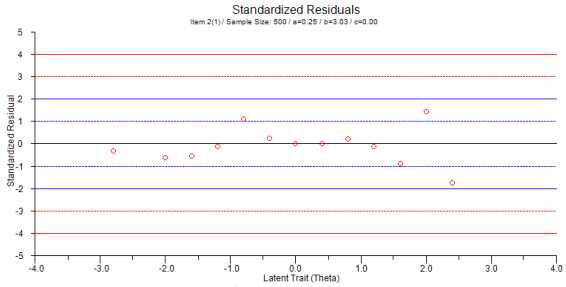

Standardized Residuals
Item 58(1) / Sample Size: 251 / a=0.24 / b=0.06 / c=0.00
Standardized Residual
Latent Trait (Theta)

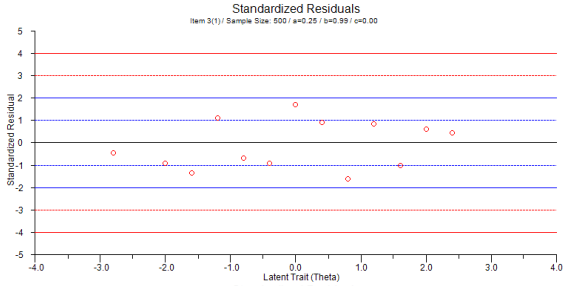

Standardized Residuals
Item 59(1) / Sample Size: 251 / a=0.24 / b=-1.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

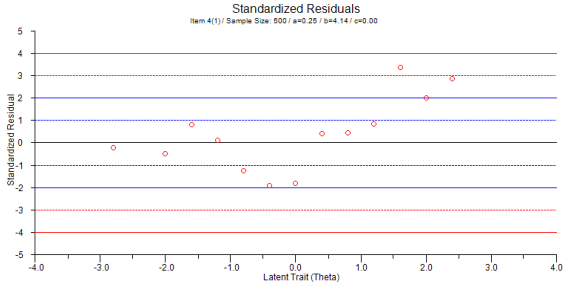

Standardized Residuals
Item 60(1) / Sample Size: 251 / a=0.24 / b=2.44 / c=0.00
Standardized Residual
Latent Trait (Theta)

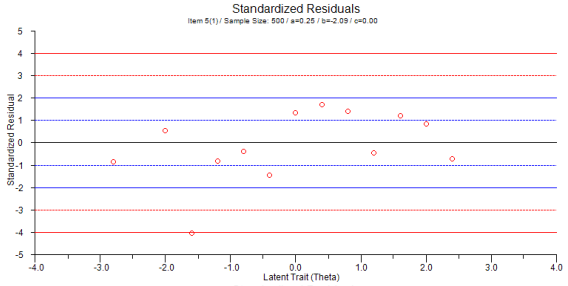
Appendix B: SRs for sample 500 1plm

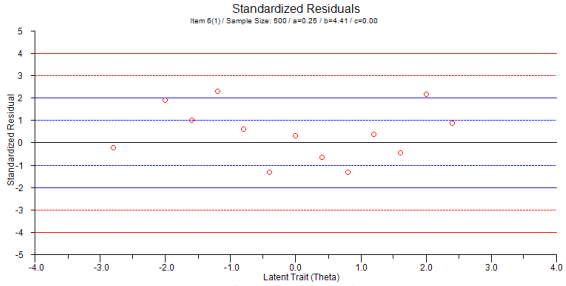


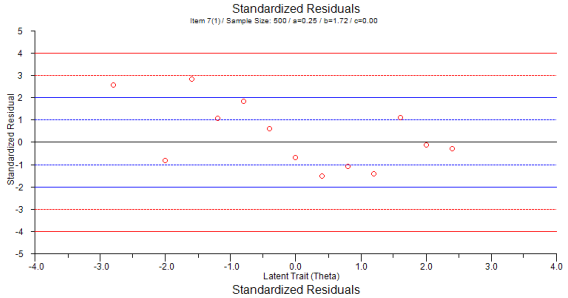


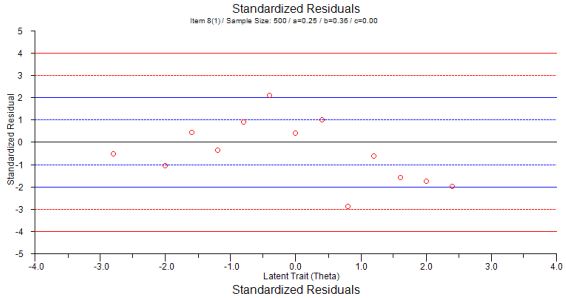


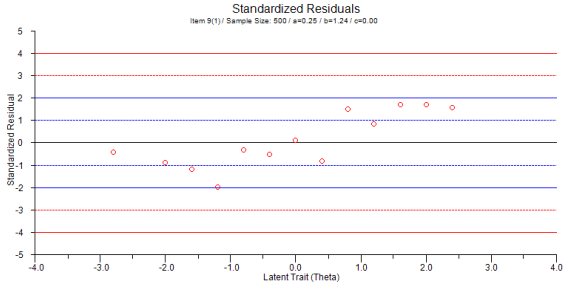


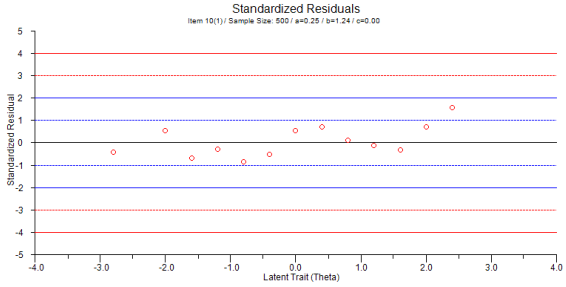


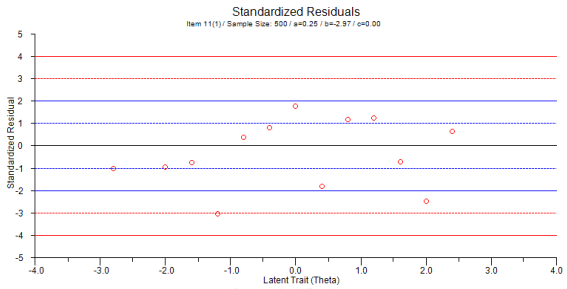


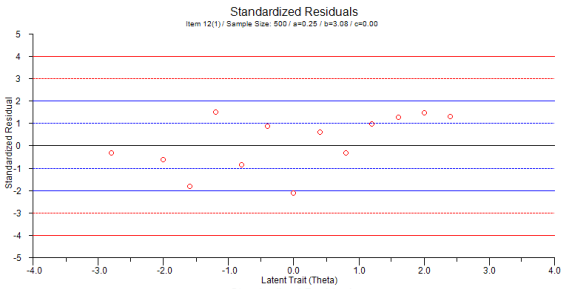


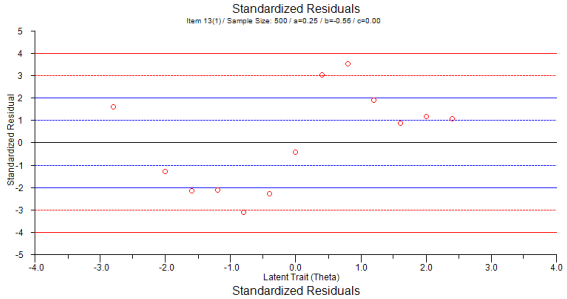


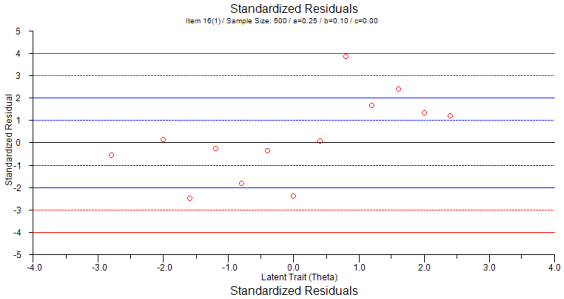


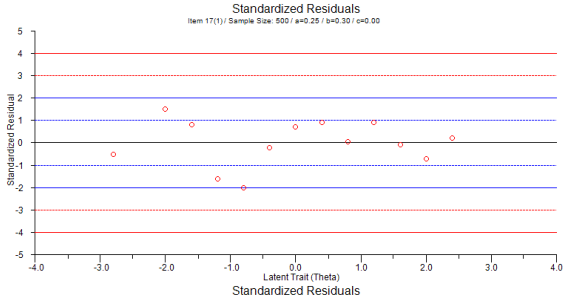


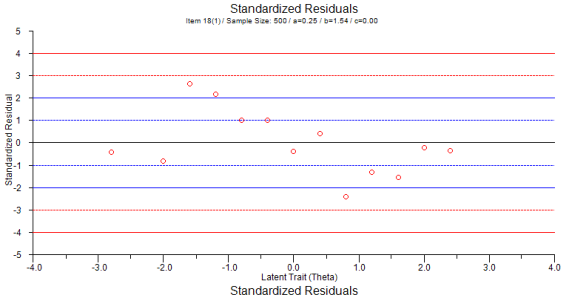

Standardized Residuals
Item 11(1) / Sample Size: 500 / a=0.25 / b=-2.97 / c=0.00
Standardized Residual
Latent Trait (Theta)

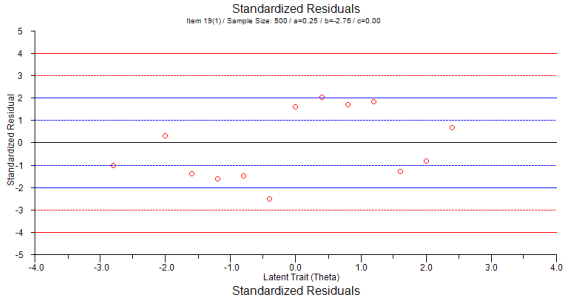

Standardized Residuals
Item 12(1) / Sample Size: 500 / a=0.25 / b=3.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

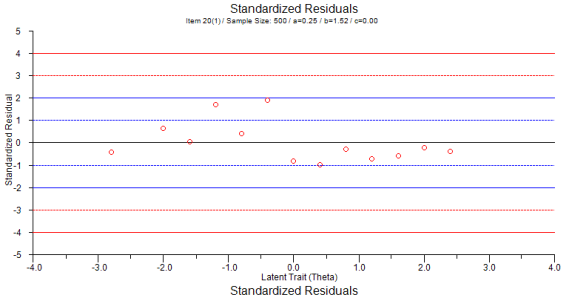

Standardized Residuals
Item 13(1) / Sample Size: 500 / a=0.25 / b=-0.56 / c=0.00
Standardized Residual
Latent Trait (Theta)

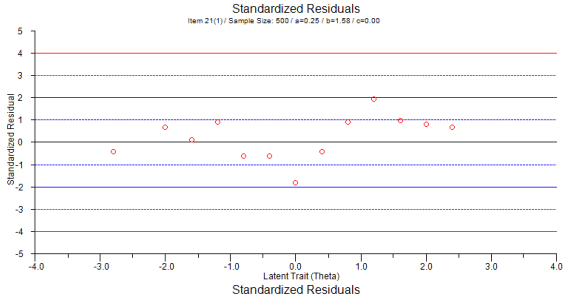

Standardized Residuals
Item 16(1) / Sample Size: 500 / a=0.25 / b=0.10 / c=0.00
Standardized Residual
Latent Trait (Theta)

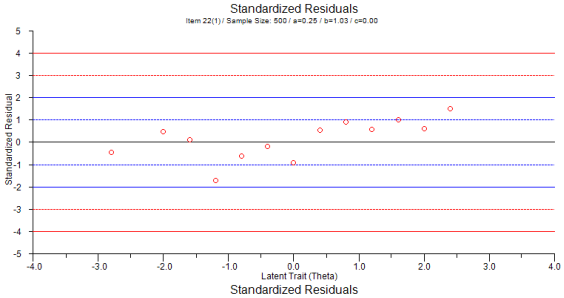

Standardized Residuals
Item 17(1) / Sample Size: 500 / a=0.25 / b=0.30 / c=0.00
Standardized Residual
Latent Trait (Theta)

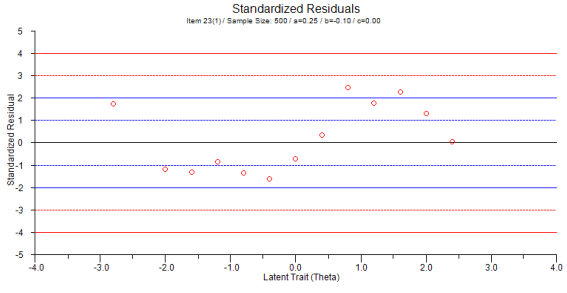

Standardized Residuals
Item 18(1) / Sample Size: 500 / a=0.25 / b=1.54 / c=0.00
Standardized Residual
Latent Trait (Theta)

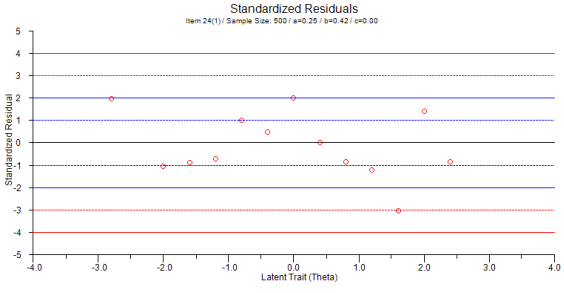

Standardized Residuals
Item 19(1) / Sample Size: 500 / a=0.25 / b=-2.76 / c=0.00
Standardized Residual
Latent Trait (Theta)

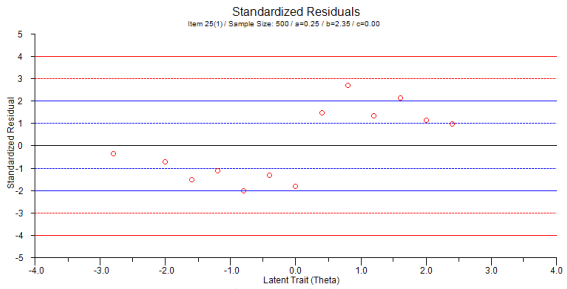

Standardized Residuals
Item 20(1) / Sample Size: 500 / a=0.25 / b=1.52 / c=0.00
Standardized Residual
Latent Trait (Theta)

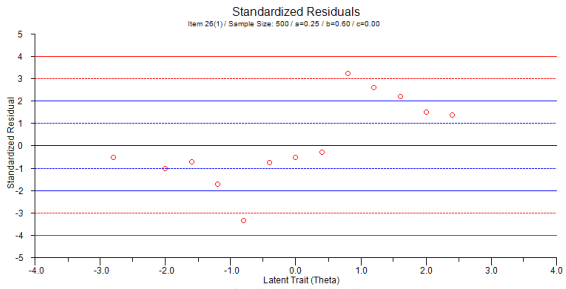

Standardized Residuals
Item 21(1) / Sample Size: 500 / a=0.25 / b=1.58 / c=0.00
Standardized Residual
Latent Trait (Theta)

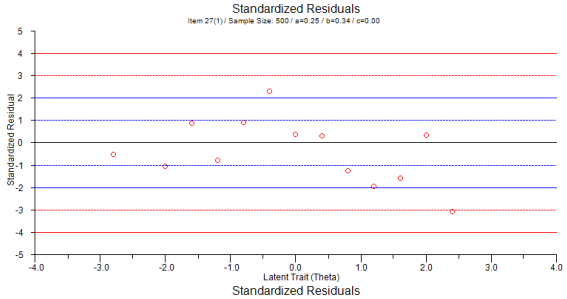

Standardized Residuals
Item 22(1) / Sample Size: 500 / a=0.25 / b=1.03 / c=0.00
Standardized Residual
Latent Trait (Theta)

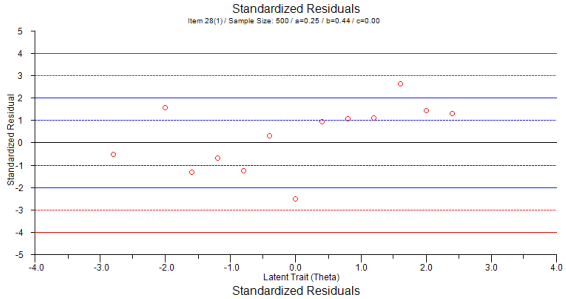

Standardized Residuals
Item 23(1) / Sample Size: 500 / a=0.25 / b=-0.10 / c=0.00
Standardized Residual
Latent Trait (Theta)

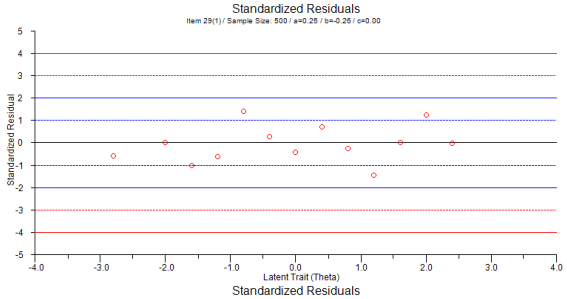

Standardized Residuals
Item 24(1) / Sample Size: 500 / a=0.25 / b=0.42 / c=0.00
Standardized Residual
Latent Trait (Theta)

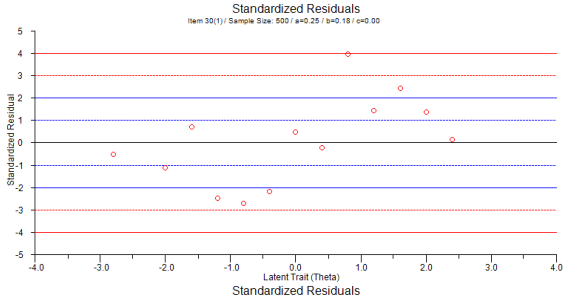

Standardized Residuals
Item 25(1) / Sample Size: 500 / a=0.25 / b=2.35 / c=0.00
Standardized Residual
Latent Trait (Theta)

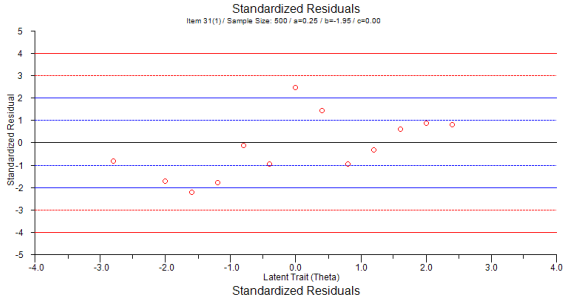

Standardized Residuals
Item 26(1) / Sample Size: 500 / a=0.25 / b=0.60 / c=0.00
Standardized Residual
Latent Trait (Theta)

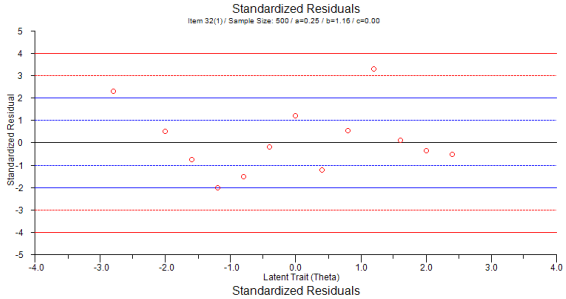

Standardized Residuals
Item 27(1) / Sample Size: 500 / a=0.25 / b=0.34 / c=0.00
Standardized Residual
Latent Trait (Theta)

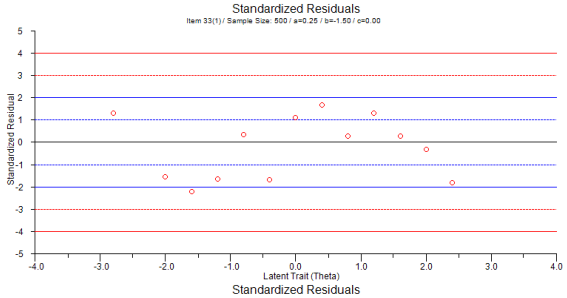

Standardized Residuals
Item 28(1) / Sample Size: 500 / a=0.25 / b=0.44 / c=0.00
Standardized Residual
Latent Trait (Theta)

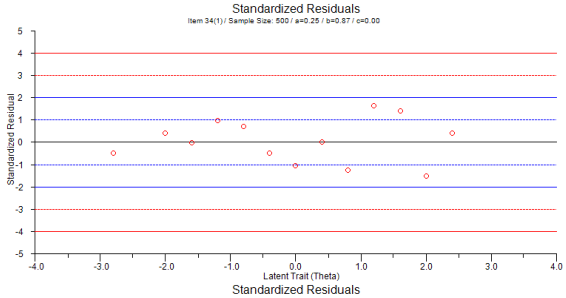

Standardized Residuals
Item 29(1) / Sample Size: 500 / a=0.25 / b=-0.26 / c=0.00
Standardized Residual
Latent Trait (Theta)

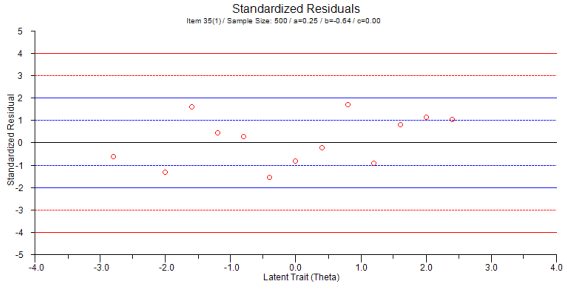

Standardized Residuals
Item 30(1) / Sample Size: 500 / a=0.25 / b=0.18 / c=0.00
Standardized Residual
Latent Trait (Theta)

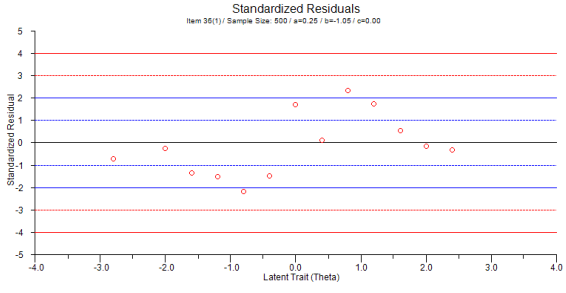

Standardized Residuals
Item 31(1) / Sample Size: 500 / a=0.25 / b=-1.95 / c=0.00
Standardized Residual
Latent Trait (Theta)

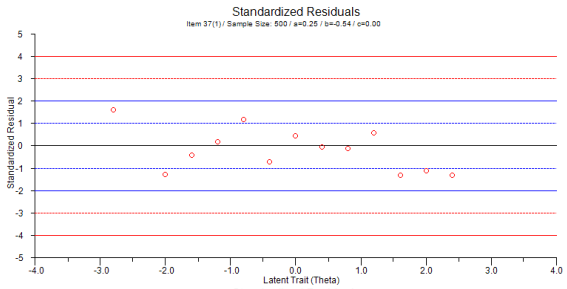

Standardized Residuals
Item 32(1) / Sample Size: 500 / a=0.25 / b=1.16 / c=0.00
Standardized Residual
Latent Trait (Theta)

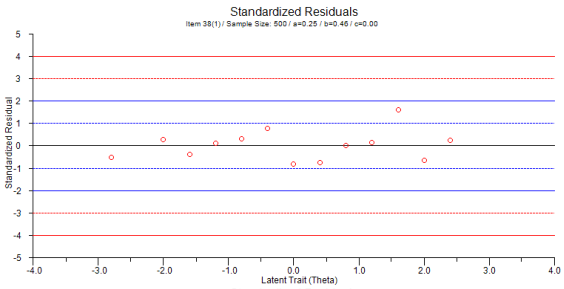

Standardized Residuals
Item 33(1) / Sample Size: 500 / a=0.25 / b=-1.50 / c=0.00
Standardized Residual
Latent Trait (Theta)

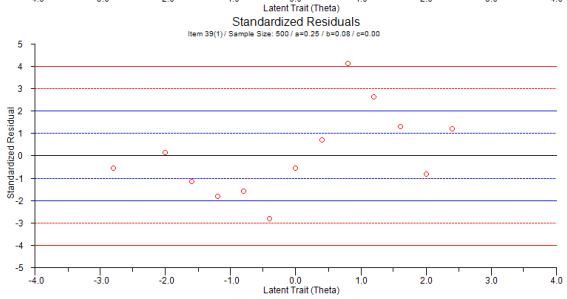

Standardized Residuals
Item 34(1) / Sample Size: 500 / a=0.25 / b=0.87 / c=0.00
Standardized Residual
Latent Trait (Theta)

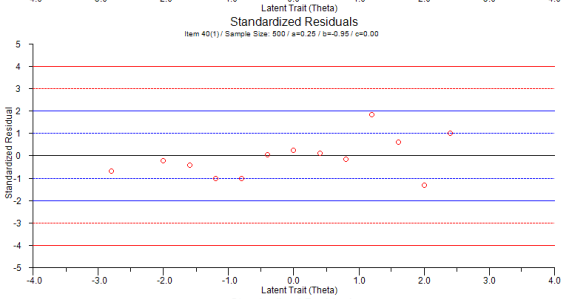

Standardized Residuals
Item 35(1) / Sample Size: 500 / a=0.25 / b=-0.64 / c=0.00
Standardized Residual
Latent Trait (Theta)

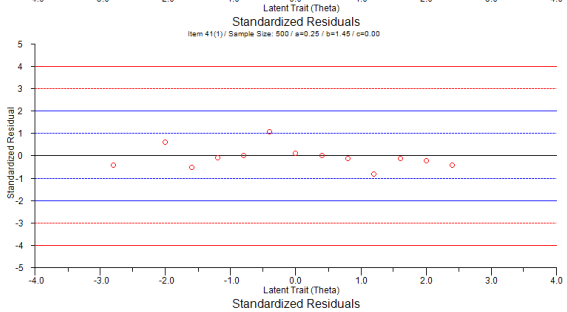

Standardized Residuals
Item 36(1) / Sample Size: 500 / a=0.25 / b=-1.05 / c=0.00
Standardized Residual
Latent Trait (Theta)

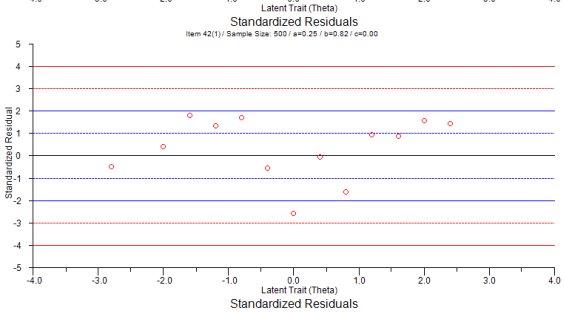

Standardized Residuals
Item 37(1) / Sample Size: 500 / a=0.25 / b=-0.54 / c=0.00
Standardized Residual
Latent Trait (Theta)

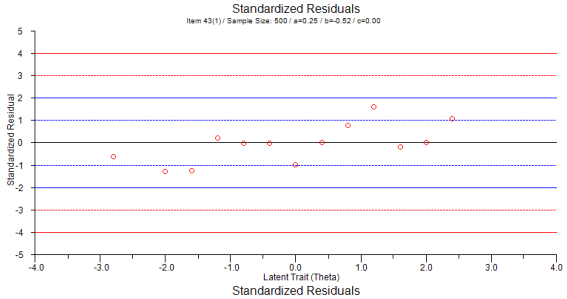

Standardized Residuals
Item 38(1) / Sample Size: 500 / a=0.25 / b=0.46 / c=0.00
Standardized Residual
Latent Trait (Theta)

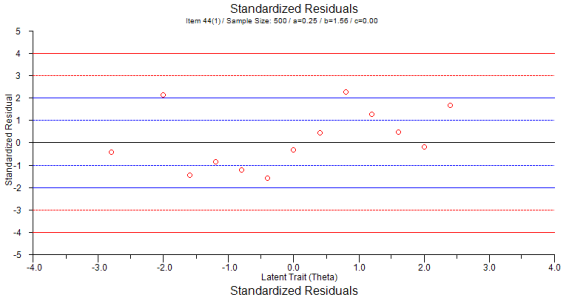

Standardized Residuals
Item 39(1) / Sample Size: 500 / a=0.25 / b=0.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

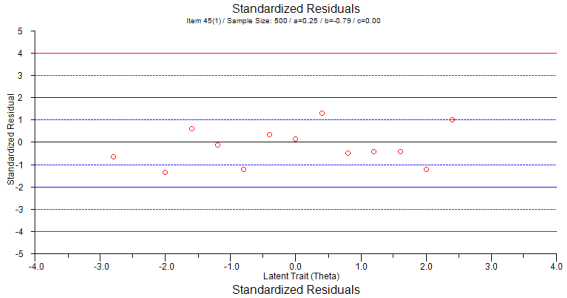

Standardized Residuals
Item 40(1) / Sample Size: 500 / a=0.25 / b=-0.95 / c=0.00
Standardized Residual
Latent Trait (Theta)

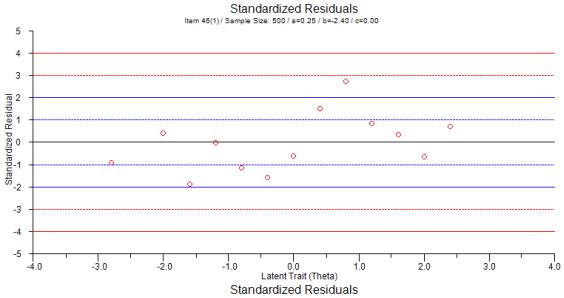

Standardized Residuals
Item 41(1) / Sample Size: 500 / a=0.25 / b=1.45 / c=0.00
Standardized Residual
Latent Trait (Theta)

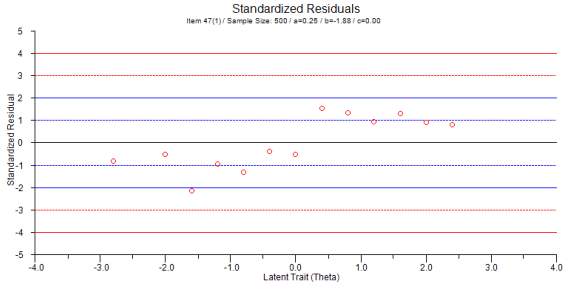

Standardized Residuals
Item 42(1) / Sample Size: 500 / a=0.25 / b=0.82 / c=0.00
Standardized Residual
Latent Trait (Theta)

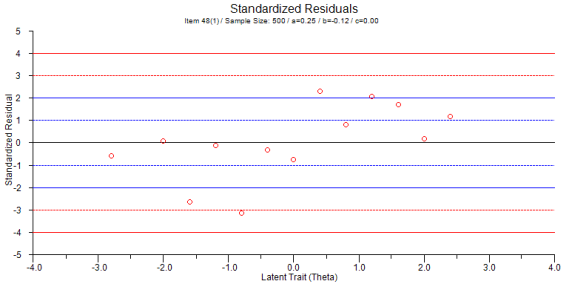

Standardized Residuals
Item 43(1) / Sample Size: 500 / a=0.25 / b=-0.52 / c=0.00
Standardized Residual
Latent Trait (Theta)

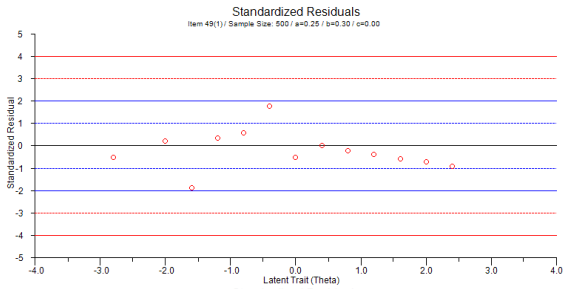

Standardized Residuals
Item 44(1) / Sample Size: 500 / a=0.25 / b=1.56 / c=0.00
Standardized Residual
Latent Trait (Theta)

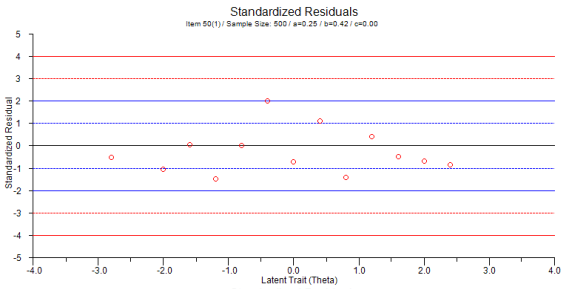

Standardized Residuals
Item 45(1) / Sample Size: 500 / a=0.25 / b=-0.79 / c=0.00
Standardized Residual
Latent Trait (Theta)

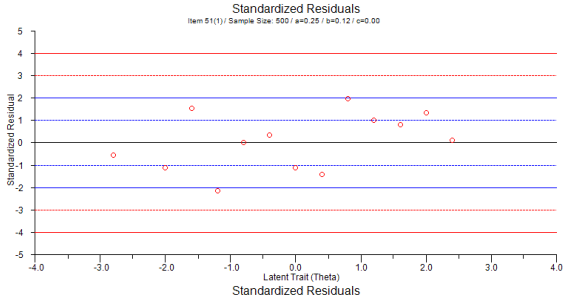

Standardized Residuals
Item 46(1) / Sample Size: 500 / a=0.25 / b=-2.40 / c=0.00
Standardized Residual
Latent Trait (Theta)

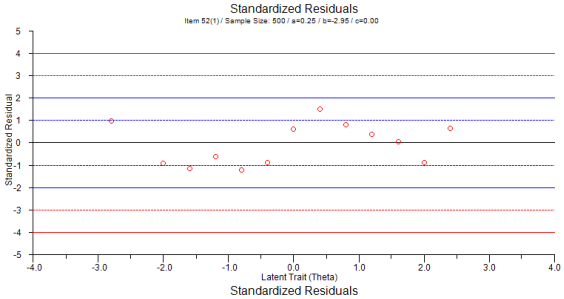

Standardized Residuals
Item 47(1) / Sample Size: 500 / a=0.25 / b=-1.88 / c=0.00
Standardized Residual
Latent Trait (Theta)

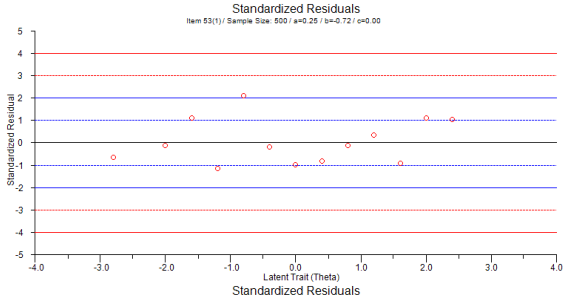

Standardized Residuals
Item 48(1) / Sample Size: 500 / a=0.25 / b=-0.12 / c=0.00
Standardized Residual
Latent Trait (Theta)

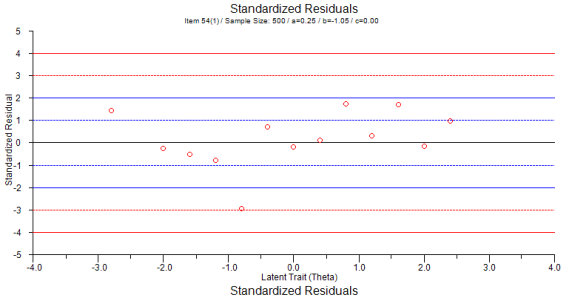

Standardized Residuals
Item 49(1) / Sample Size: 500 / a=0.25 / b=0.30 / c=0.00
Standardized Residual
Latent Trait (Theta)

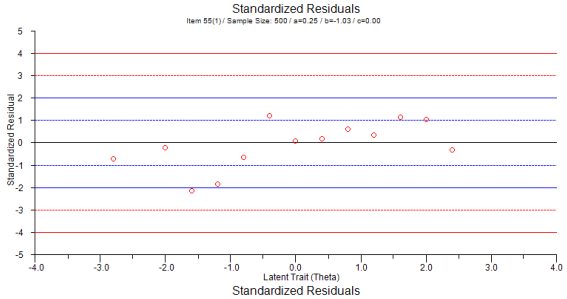

Standardized Residuals
Item 50(1) / Sample Size: 500 / a=0.25 / b=0.42 / c=0.00
Standardized Residual
Latent Trait (Theta)

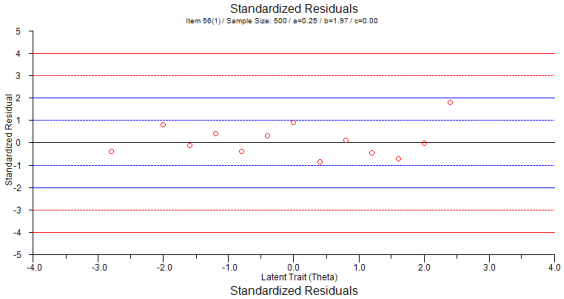

Standardized Residuals
Item 51(1) / Sample Size: 500 / a=0.25 / b=0.12 / c=0.00
Standardized Residual
Latent Trait (Theta)

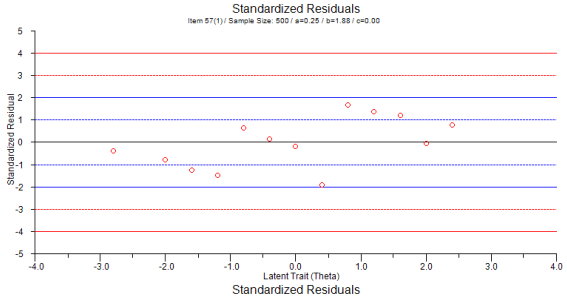

Standardized Residuals
Item 52(1) / Sample Size: 500 / a=0.25 / b=-2.95 / c=0.00
Standardized Residual
Latent Trait (Theta)

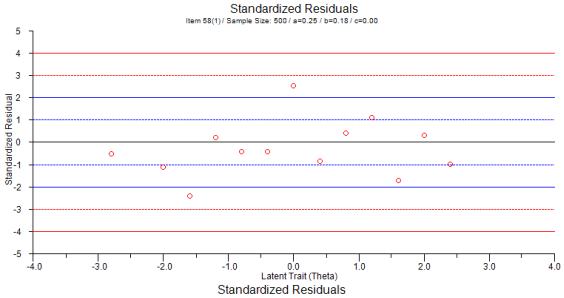

Standardized Residuals
Item 53(1) / Sample Size: 500 / a=0.25 / b=-0.72 / c=0.00
Standardized Residual
Latent Trait (Theta)

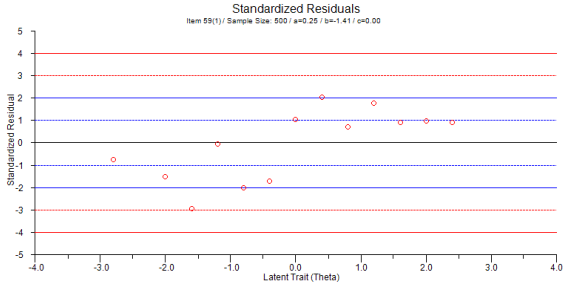

Standardized Residuals
Item 54(1) / Sample Size: 500 / a=0.25 / b=-1.05 / c=0.00
Standardized Residual
Latent Trait (Theta)

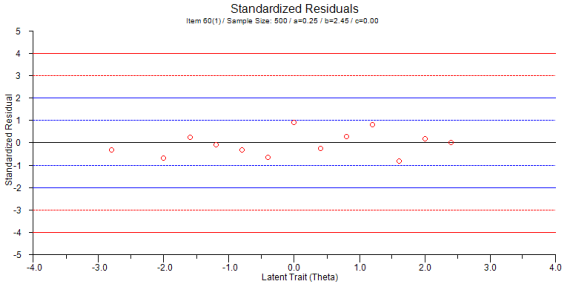

Standardized Residuals
Item 55(1) / Sample Size: 500 / a=0.25 / b=-1.03 / c=0.00
Standardized Residual
Latent Trait (Theta)


Standardized Residuals
Item 56(1) / Sample Size: 500 / a=0.25 / b=1.97 / c=0.00
Standardized Residual
Latent Trait (Theta)

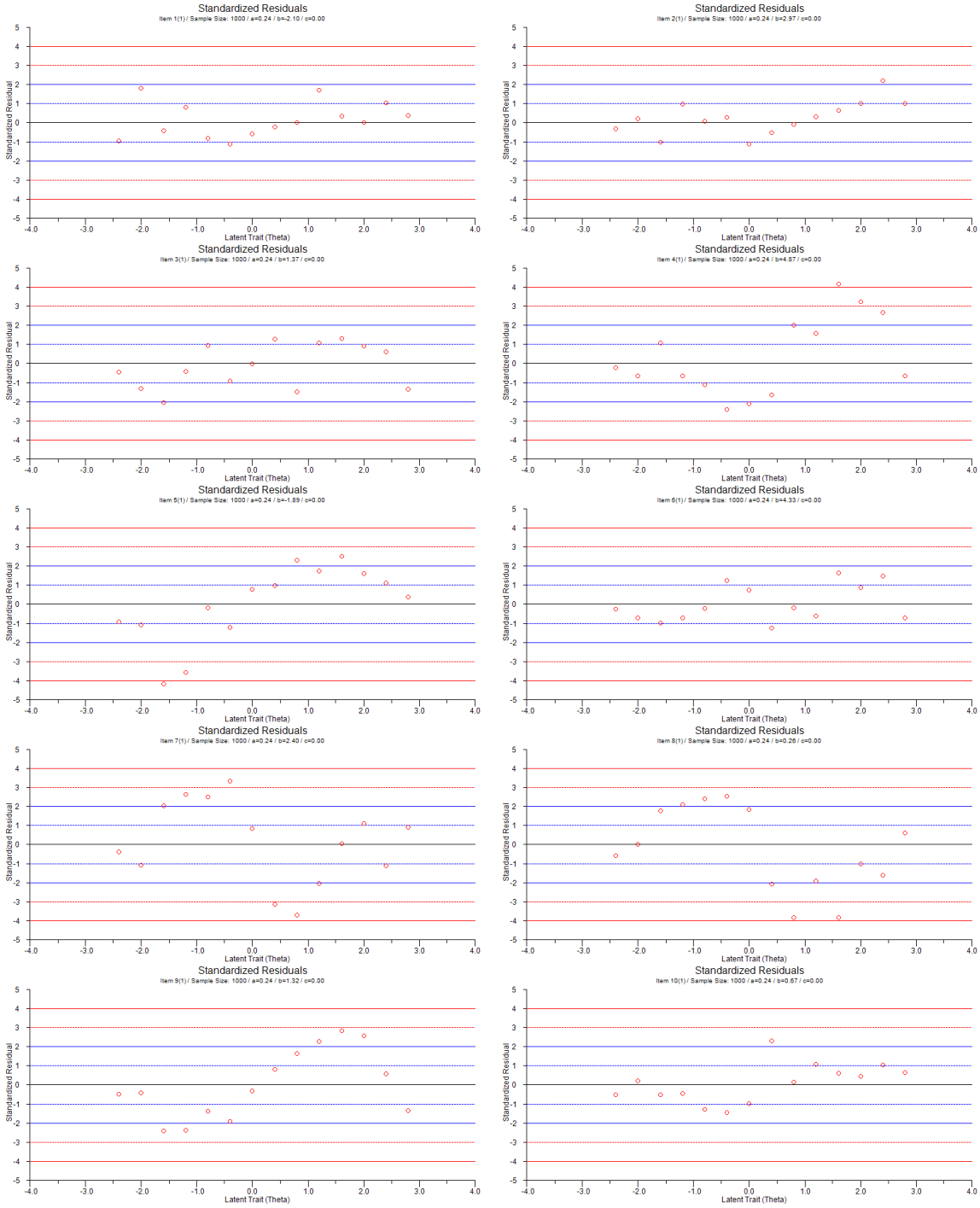

Standardized Residuals
Item 57(1) / Sample Size: 500 / a=0.25 / b=1.88 / c=0.00
Standardized Residual
Latent Trait (Theta)

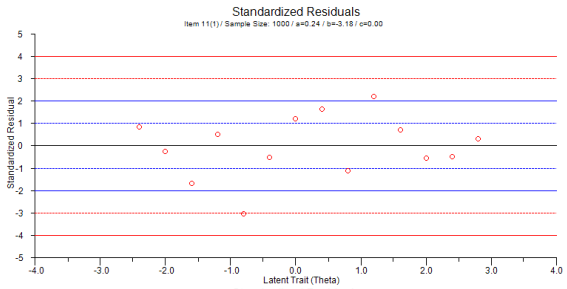

Standardized Residuals
Item 58(1) / Sample Size: 500 / a=0.25 / b=0.18 / c=0.00
Standardized Residual
Latent Trait (Theta)

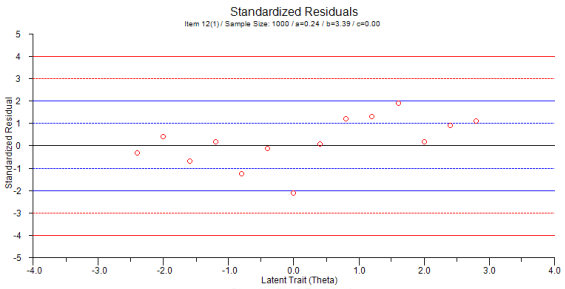

Standardized Residuals
Item 59(1) / Sample Size: 500 / a=0.25 / b=-1.41 / c=0.00
Standardized Residual
Latent Trait (Theta)

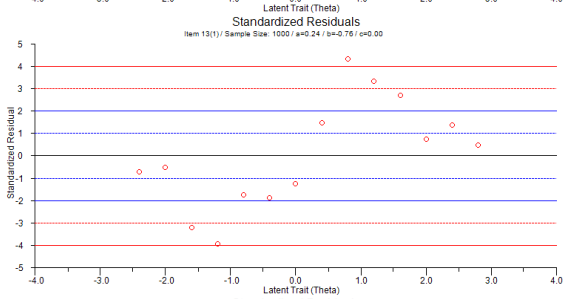

Standardized Residuals
Item 60(1) / Sample Size: 500 / a=0.25 / b=2.45 / c=0.00
Standardized Residual
Latent Trait (Theta)

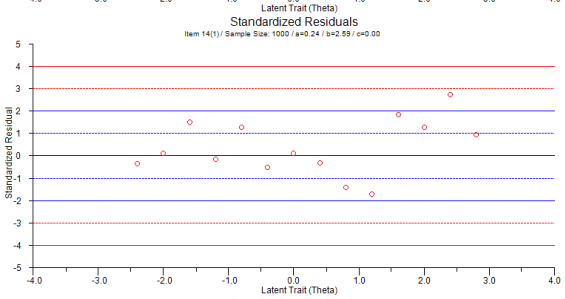
Appendix C: SR s for 1plm sample 1000

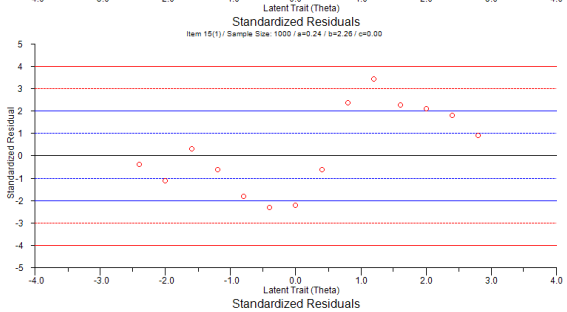


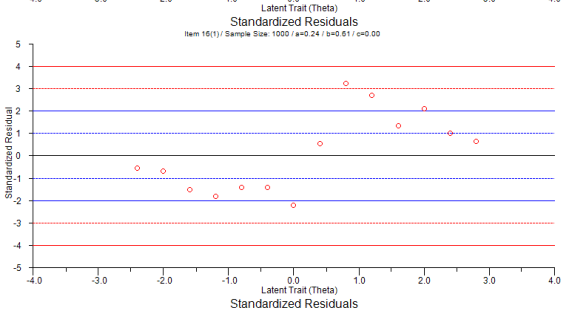

Standardized Residuals
Item 11(1) / Sample Size: 1000 / a=0.24 / b=-3.18 / c=0.00
Standardized Residual
Latent Trait (Theta)

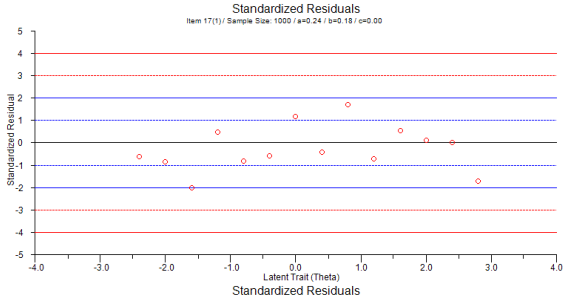

Standardized Residuals
Item 12(1) / Sample Size: 1000 / a=0.24 / b=3.39 / c=0.00
Standardized Residual
Latent Trait (Theta)

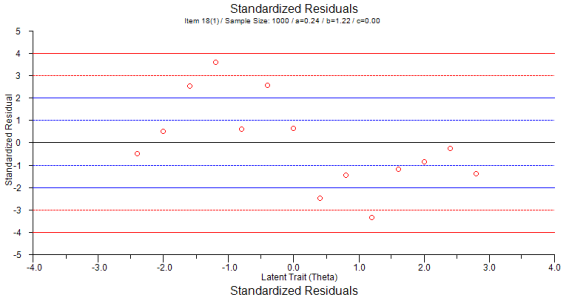

Standardized Residuals
Item 13(1) / Sample Size: 1000 / a=0.24 / b=-0.76 / c=0.00
Standardized Residual
Latent Trait (Theta)

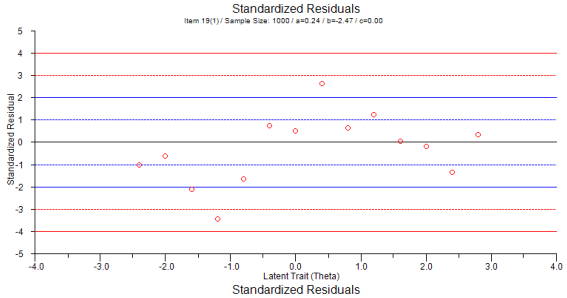

Standardized Residuals
Item 14(1) / Sample Size: 1000 / a=0.24 / b=2.59 / c=0.00
Standardized Residual
Latent Trait (Theta)

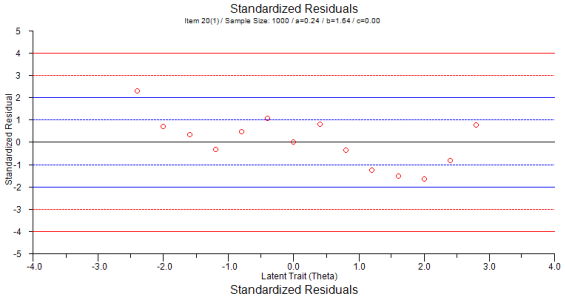

Standardized Residuals
Item 15(1) / Sample Size: 1000 / a=0.24 / b=2.26 / c=0.00
Standardized Residual
Latent Trait (Theta)

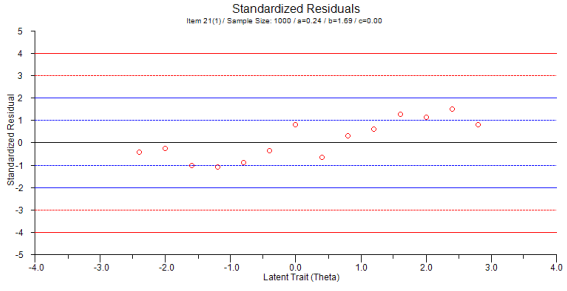

Standardized Residuals
Item 16(1) / Sample Size: 1000 / a=0.24 / b=0.61 / c=0.00
Standardized Residual
Latent Trait (Theta)

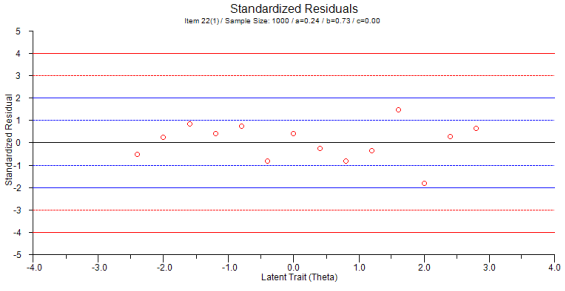

Standardized Residuals
Item 17(1) / Sample Size: 1000 / a=0.24 / b=0.18 / c=0.00
Standardized Residual
Latent Trait (Theta)

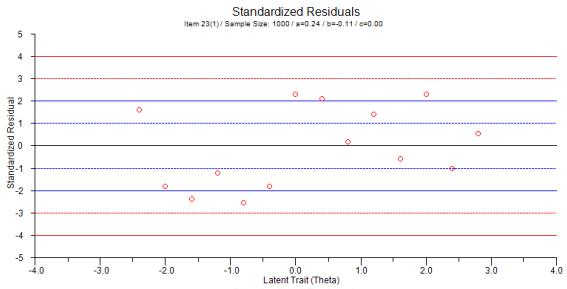

Standardized Residuals
Item 18(1) / Sample Size: 1000 / a=0.24 / b=1.22 / c=0.00
Standardized Residual
Latent Trait (Theta)

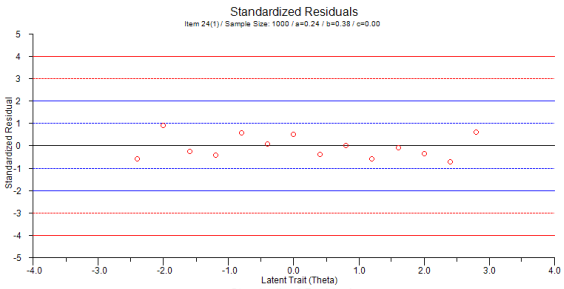

Standardized Residuals
Item 19(1) / Sample Size: 1000 / a=0.24 / b=-2.47 / c=0.00
Standardized Residual
Latent Trait (Theta)

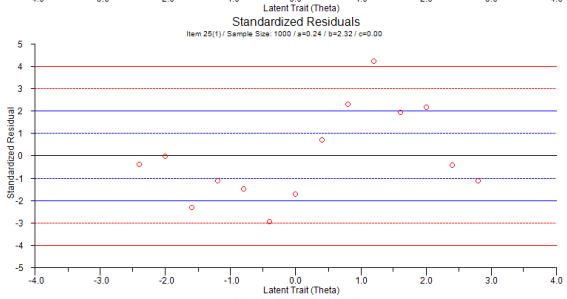

Standardized Residuals
Item 20(1) / Sample Size: 1000 / a=0.24 / b=1.64 / c=0.00
Standardized Residual
Latent Trait (Theta)

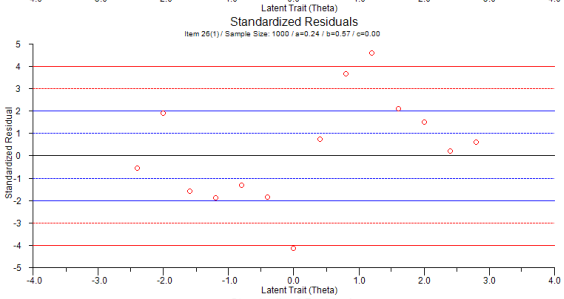

Standardized Residuals
Item 21(1) / Sample Size: 1000 / a=0.24 / b=1.69 / c=0.00
Standardized Residual
Latent Trait (Theta)

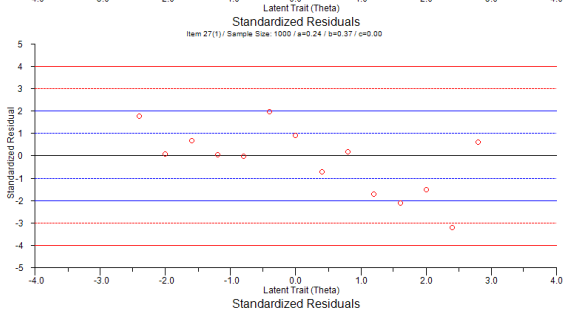

Standardized Residuals
Item 22(1) / Sample Size: 1000 / a=0.24 / b=0.73 / c=0.00
Standardized Residual
Latent Trait (Theta)

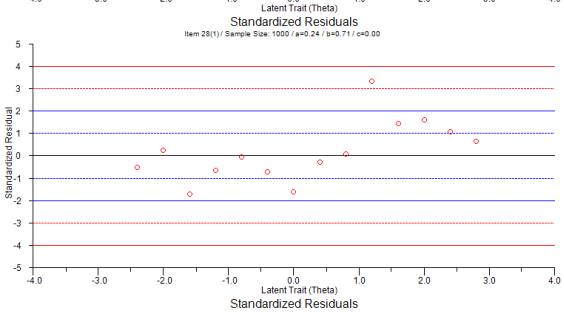

Standardized Residuals
Item 23(1) / Sample Size: 1000 / a=0.24 / b=-0.11 / c=0.00
Standardized Residual
Latent Trait (Theta)

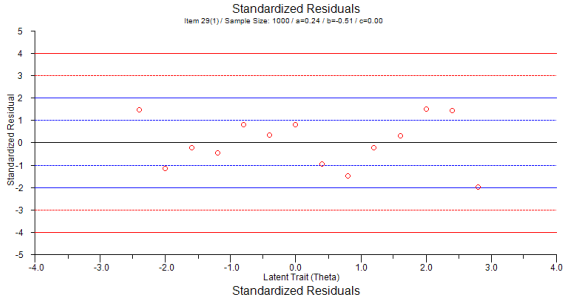

Standardized Residuals
Item 24(1) / Sample Size: 1000 / a=0.24 / b=0.38 / c=0.00
Standardized Residual
Latent Trait (Theta)

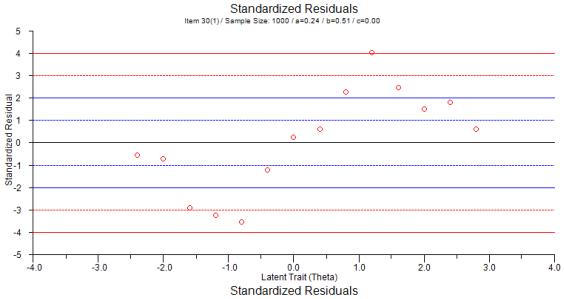

Standardized Residuals
Item 25(1) / Sample Size: 1000 / a=0.24 / b=2.32 / c=0.00
Standardized Residual
Latent Trait (Theta)

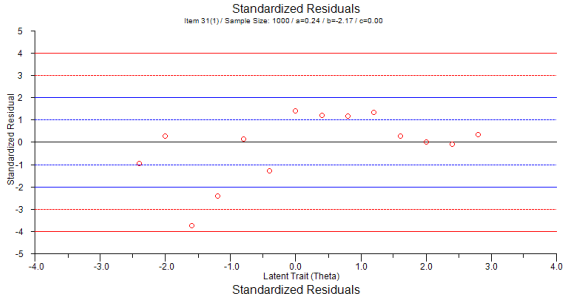

Standardized Residuals
Item 26(1) / Sample Size: 1000 / a=0.24 / b=0.57 / c=0.00
Standardized Residual
Latent Trait (Theta)

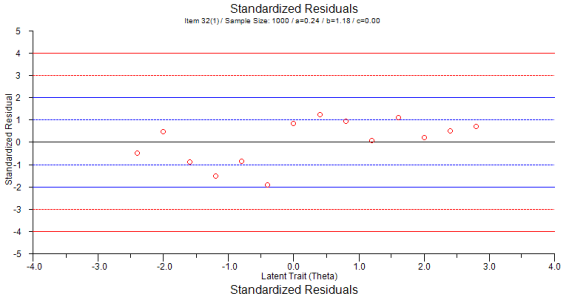

Standardized Residuals
Item 27(1) / Sample Size: 1000 / a=0.24 / b=0.37 / c=0.00
Standardized Residual
Latent Trait (Theta)

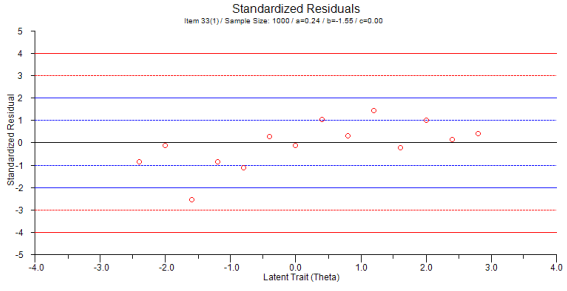

Standardized Residuals
Item 28(1) / Sample Size: 1000 / a=0.24 / b=0.71 / c=0.00
Standardized Residual
Latent Trait (Theta)

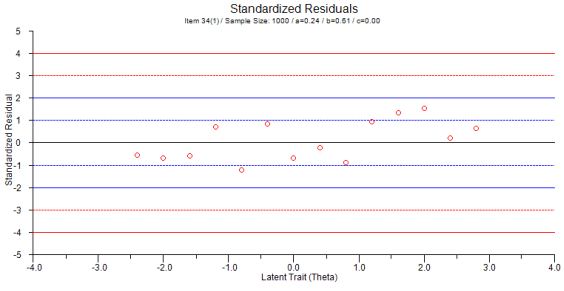

Standardized Residuals
Item 29(1) / Sample Size: 1000 / a=0.24 / b=-0.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

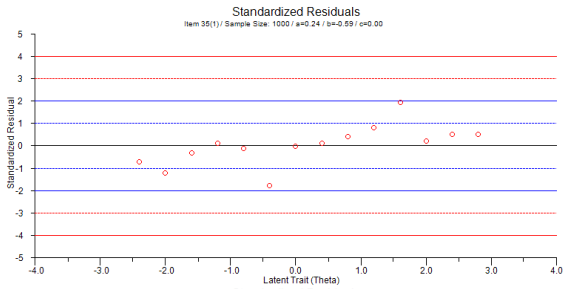

Standardized Residuals
Item 30(1) / Sample Size: 1000 / a=0.24 / b=0.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

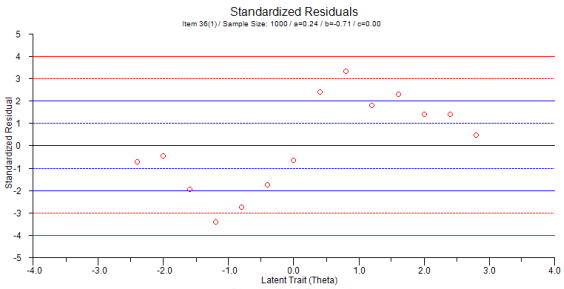

Standardized Residuals
Item 31(1) / Sample Size: 1000 / a=0.24 / b=-2.17 / c=0.00
Standardized Residual
Latent Trait (Theta)

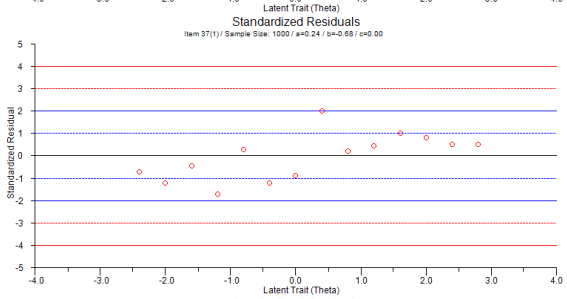

Standardized Residuals
Item 32(1) / Sample Size: 1000 / a=0.24 / b=1.18 / c=0.00
Standardized Residual
Latent Trait (Theta)

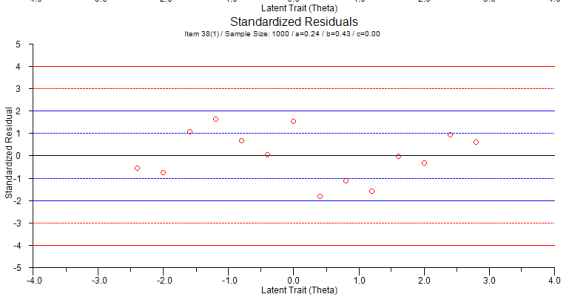

Standardized Residuals
Item 33(1) / Sample Size: 1000 / a=0.24 / b=-1.55 / c=0.00
Standardized Residual
Latent Trait (Theta)

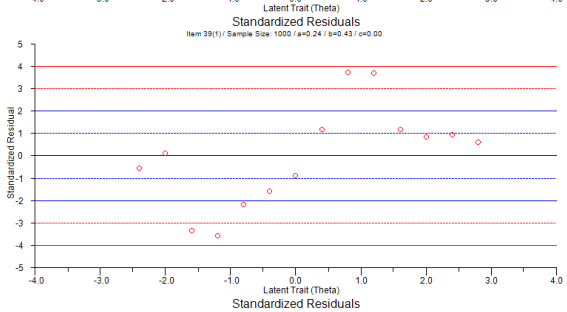

Standardized Residuals
Item 34(1) / Sample Size: 1000 / a=0.24 / b=0.61 / c=0.00
Standardized Residual
Latent Trait (Theta)

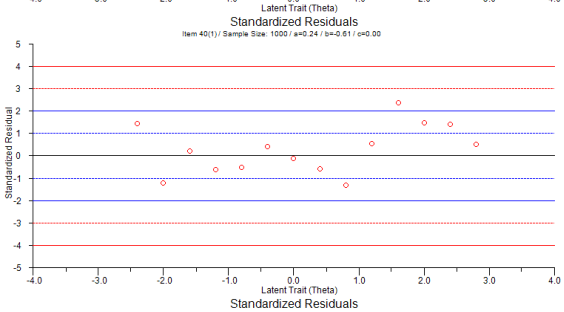

Standardized Residuals
Item 35(1) / Sample Size: 1000 / a=0.24 / b=-0.59 / c=0.00
Standardized Residual
Latent Trait (Theta)

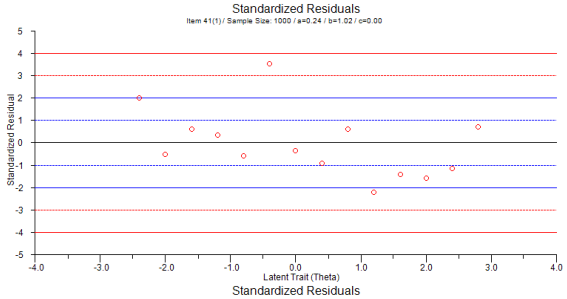

Standardized Residuals
Item 36(1) / Sample Size: 1000 / a=0.24 / b=-0.71 / c=0.00
Standardized Residual
Latent Trait (Theta)

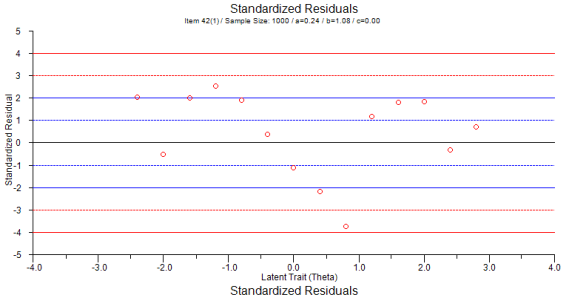

Standardized Residuals
Item 37(1) / Sample Size: 1000 / a=0.24 / b=-0.68 / c=0.00
Standardized Residual
Latent Trait (Theta)

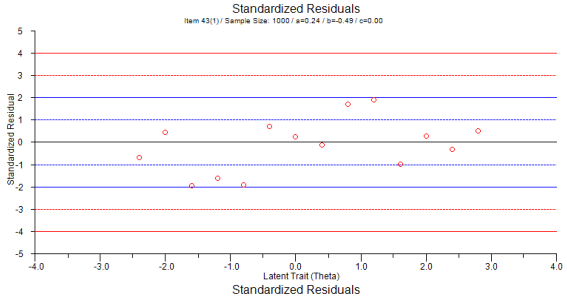

Standardized Residuals
Item 38(1) / Sample Size: 1000 / a=0.24 / b=0.43 / c=0.00
Standardized Residual
Latent Trait (Theta)

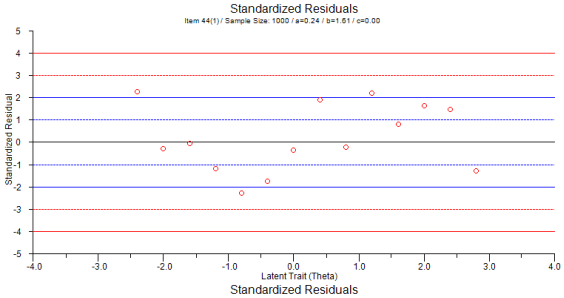

Standardized Residuals
Item 39(1) / Sample Size: 1000 / a=0.24 / b=0.43 / c=0.00
Standardized Residual
Latent Trait (Theta)

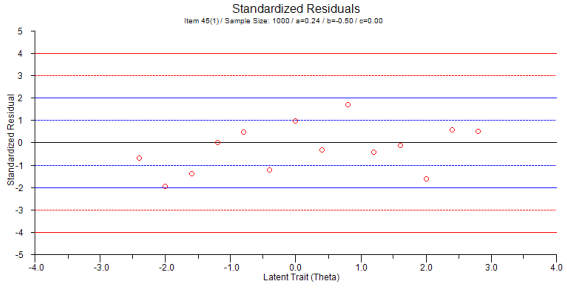

Standardized Residuals
Item 40(1) / Sample Size: 1000 / a=0.24 / b=-0.61 / c=0.00
Standardized Residual
Latent Trait (Theta)

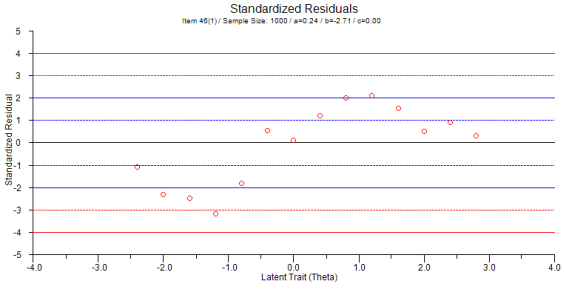

Standardized Residuals
Item 41(1) / Sample Size: 1000 / a=0.24 / b=1.02 / c=0.00
Standardized Residual
Latent Trait (Theta)

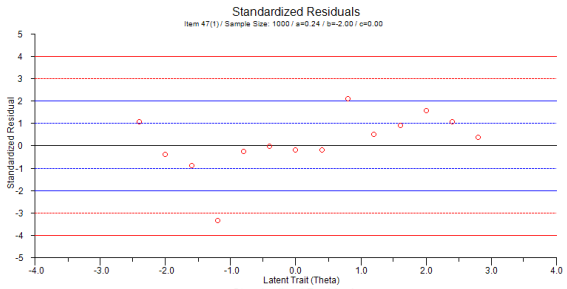

Standardized Residuals
Item 42(1) / Sample Size: 1000 / a=0.24 / b=1.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

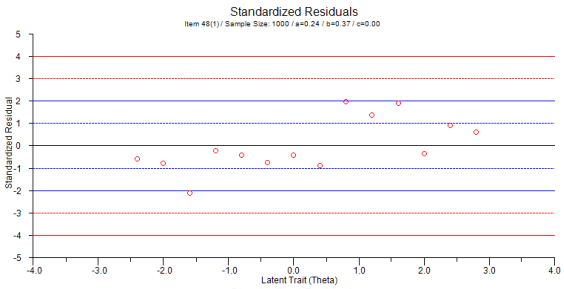

Standardized Residuals
Item 43(1) / Sample Size: 1000 / a=0.24 / b=-0.49 / c=0.00
Standardized Residual
Latent Trait (Theta)

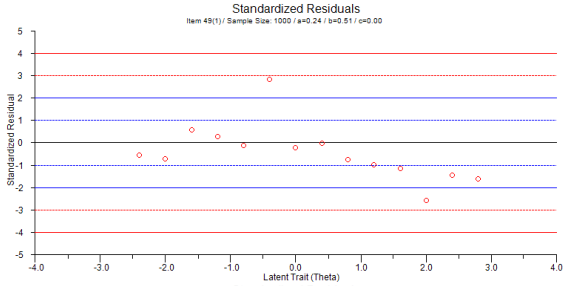

Standardized Residuals
Item 44(1) / Sample Size: 1000 / a=0.24 / b=1.61 / c=0.00
Standardized Residual
Latent Trait (Theta)

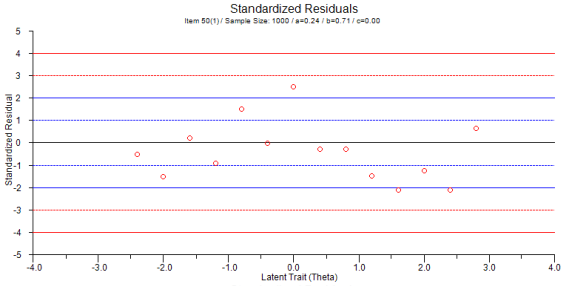

Standardized Residuals
Item 45(1) / Sample Size: 1000 / a=0.24 / b=-0.50 / c=0.00
Standardized Residual
Latent Trait (Theta)

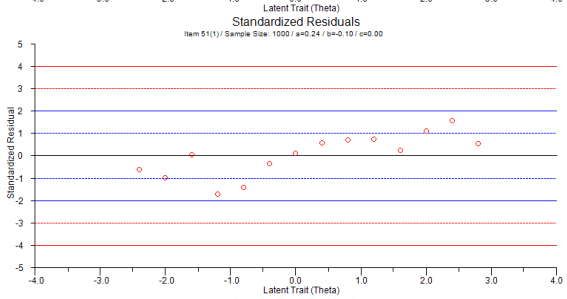

Standardized Residuals
Item 46(1) / Sample Size: 1000 / a=0.24 / b=2.71 / c=0.00
Standardized Residual
Latent Trait (Theta)

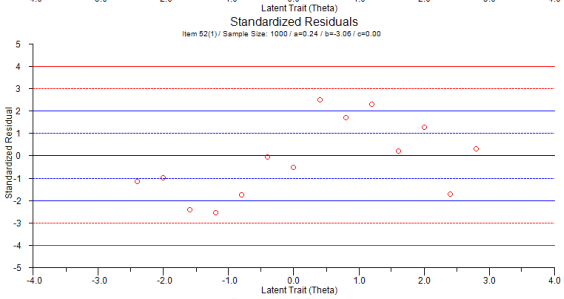

Standardized Residuals
Item 47(1) / Sample Size: 1000 / a=0.24 / b=-2.00 / c=0.00
Standardized Residual
Latent Trait (Theta)

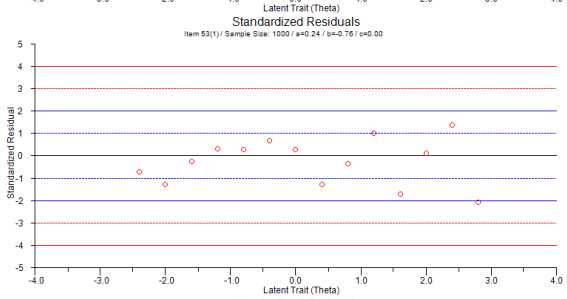

Standardized Residuals
Item 48(1) / Sample Size: 1000 / a=0.24 / b=0.37 / c=0.00
Standardized Residual
Latent Trait (Theta)

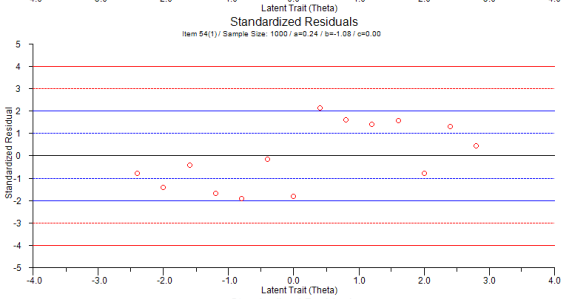

Standardized Residuals
Item 49(1) / Sample Size: 1000 / a=0.24 / b=0.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

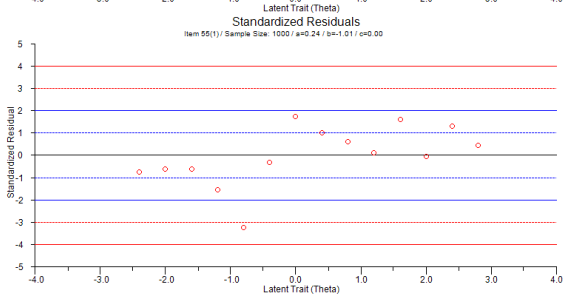

Standardized Residuals
Item 50(1) / Sample Size: 1000 / a=0.24 / b=0.71 / c=0.00
Standardized Residual
Latent Trait (Theta)

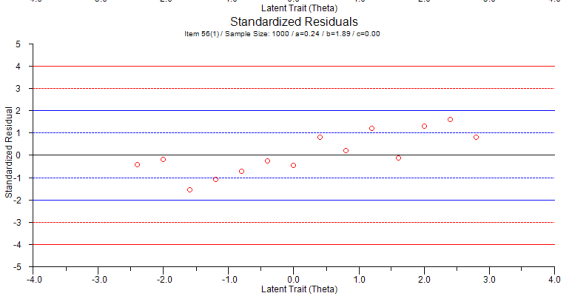

Standardized Residuals
Item 51(1) / Sample Size: 1000 / a=0.24 / b=-0.10 / c=0.00
Standardized Residual
Latent Trait (Theta)

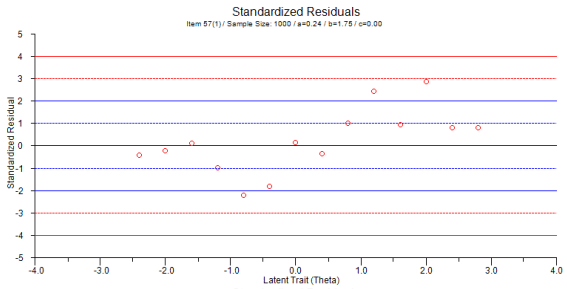

Standardized Residuals
Item 52(1) / Sample Size: 1000 / a=0.24 / b=-3.05 / c=0.00
Standardized Residual
Latent Trait (Theta)

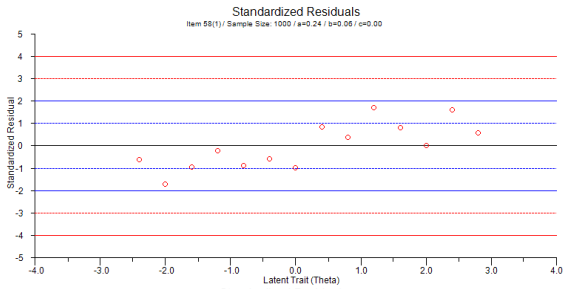

Standardized Residuals
Item 53(1) / Sample Size: 1000 / a=0.24 / b=-0.76 / c=0.00
Standardized Residual
Latent Trait (Theta)

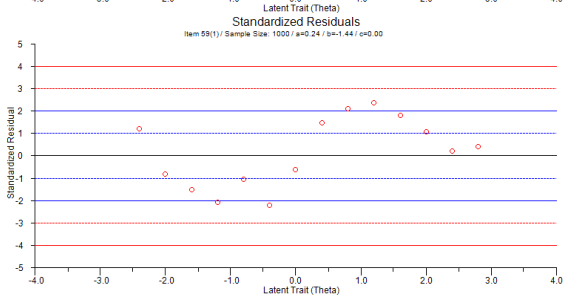

Standardized Residuals
Item 54(1) / Sample Size: 1000 / a=0.24 / b=-1.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

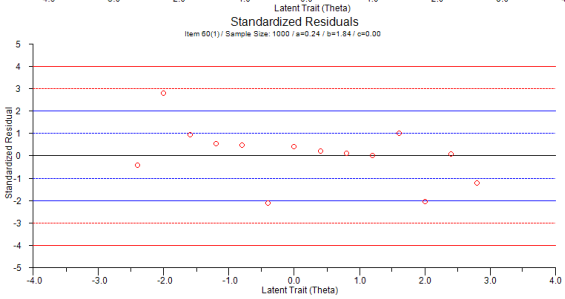

Standardized Residuals
Item 55(1) / Sample Size: 1000 / a=0.24 / b=-1.01 / c=0.00
Standardized Residual
Latent Trait (Theta)


Standardized Residuals
Item 56(1) / Sample Size: 1000 / a=0.24 / b=1.89 / c=0.00
Standardized Residual
Latent Trait (Theta)

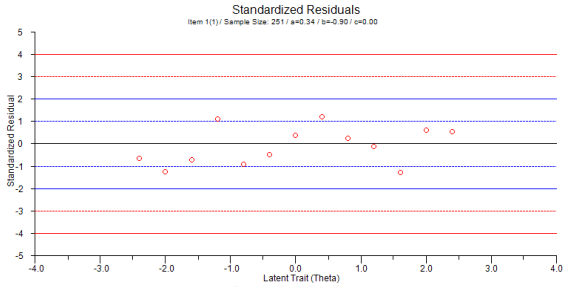

Standardized Residuals
Item 57(1) / Sample Size: 1000 / a=0.24 / b=1.75 / c=0.00
Standardized Residual
Latent Trait (Theta)

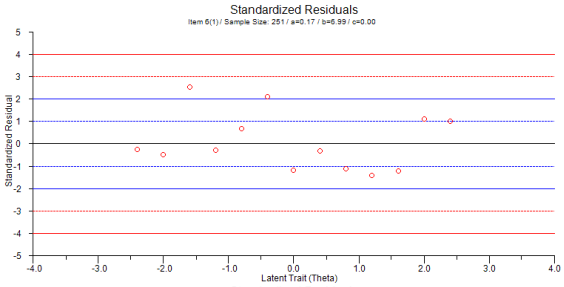

Standardized Residuals
Item 58(1) / Sample Size: 1000 / a=0.24 / b=0.06 / c=0.00
Standardized Residual
Latent Trait (Theta)

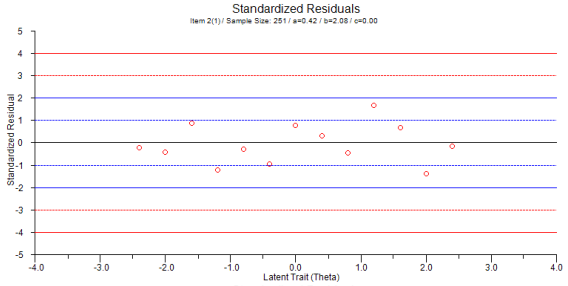

Standardized Residuals
Item 59(1) / Sample Size: 1000 / a=0.24 / b=-1.44 / c=0.00
Standardized Residual
Latent Trait (Theta)

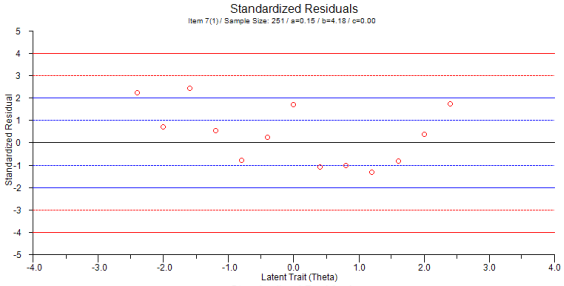

Standardized Residuals
Item 60(1) / Sample Size: 1000 / a=0.24 / b=1.84 / c=0.00
Standardized Residual
Latent Trait (Theta)

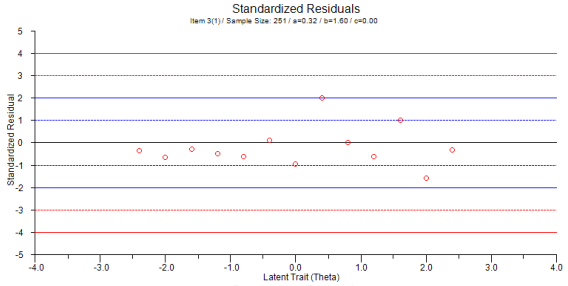
Appendix D: SRs 2PLM for sample 250

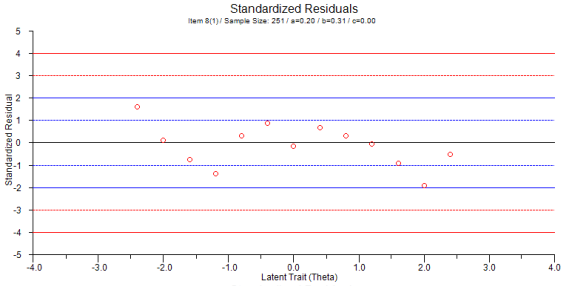


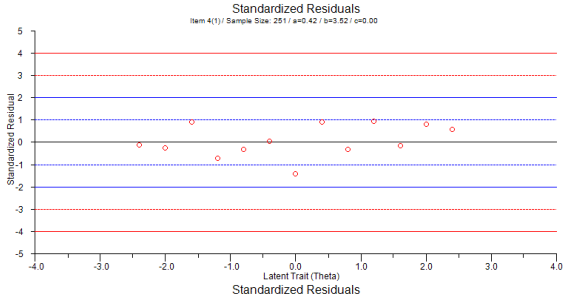


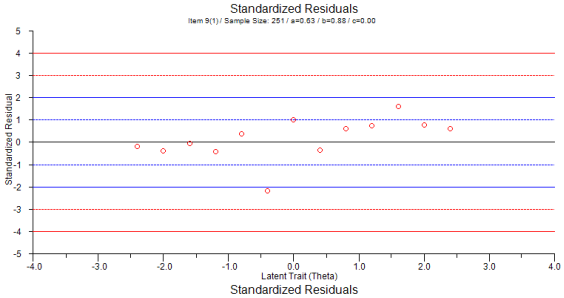


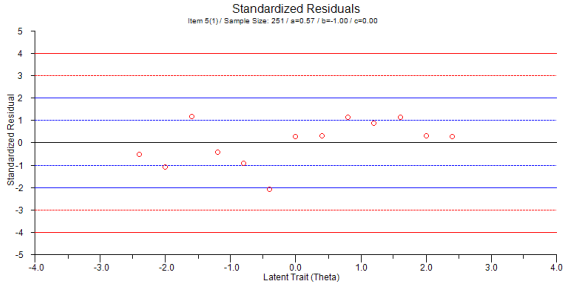


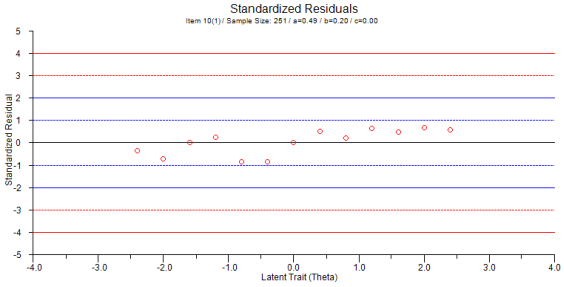


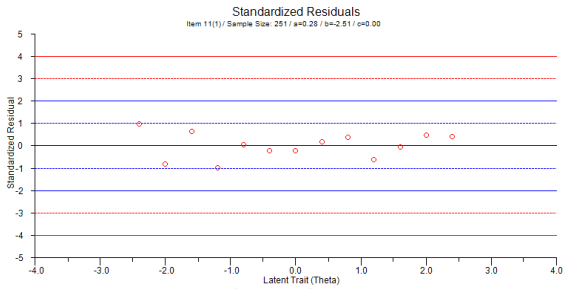


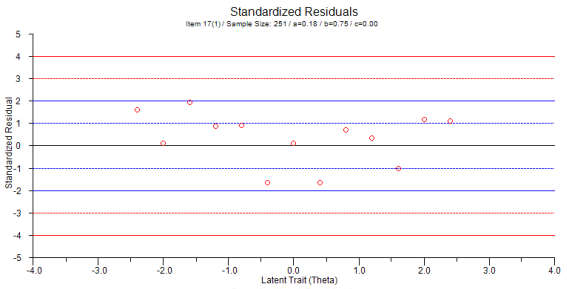


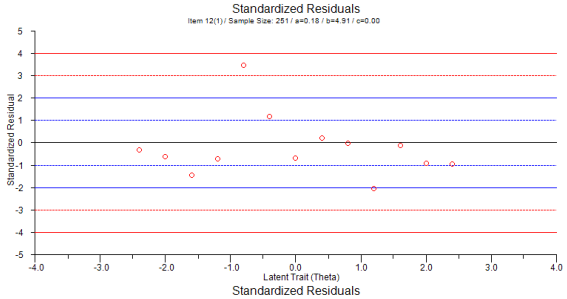


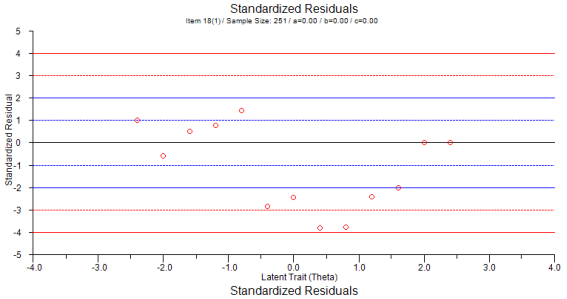


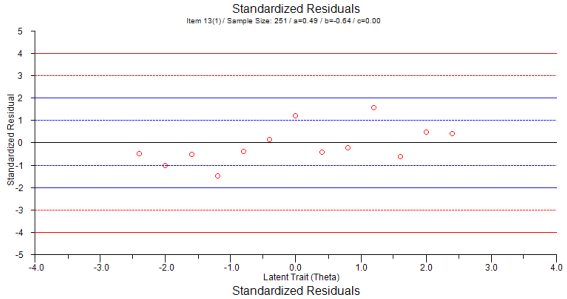


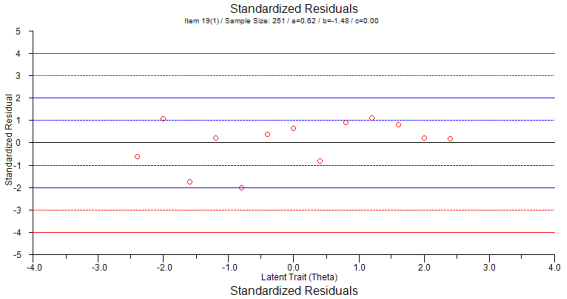

Standardized Residuals
Item 11(1) / Sample Size: 251 / a=0.28 / b=-2.51 / c=0.00
Standardized Residual
Latent Trait (Theta)

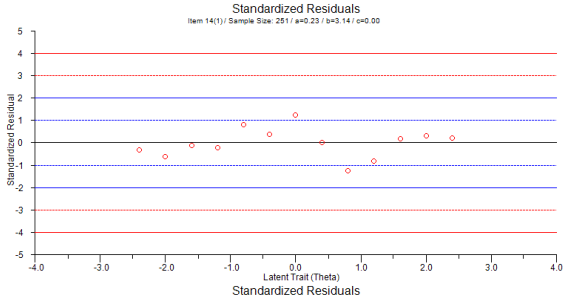

Standardized Residuals
Item 17(1) / Sample Size: 251 / a=0.18 / b=0.75 / c=0.00
Standardized Residual
Latent Trait (Theta)

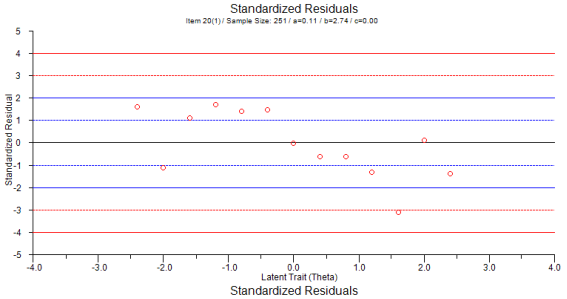

Standardized Residuals
Item 12(1) / Sample Size: 251 / a=0.18 / b=4.91 / c=0.00
Standardized Residual
Latent Trait (Theta)

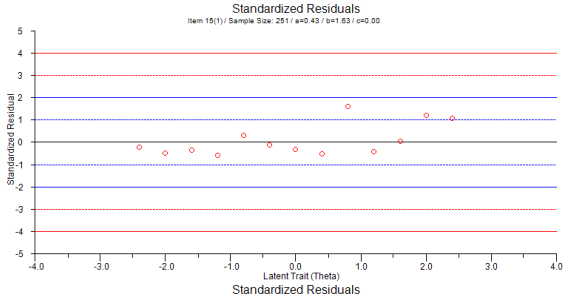

Standardized Residuals
Item 18(1) / Sample Size: 251 / a=0.00 / b=0.00 / c=0.00
Standardized Residual
Latent Trait (Theta)

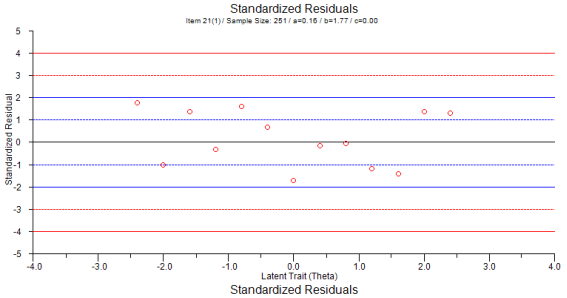

Standardized Residuals
Item 13(1) / Sample Size: 251 / a=0.49 / b=-0.64 / c=0.00
Standardized Residual
Latent Trait (Theta)

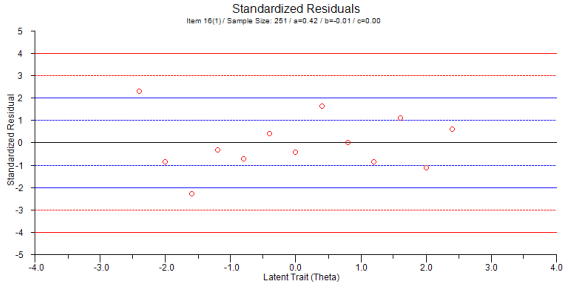

Standardized Residuals
Item 19(1) / Sample Size: 251 / a=0.62 / b=-1.48 / c=0.00
Standardized Residual
Latent Trait (Theta)

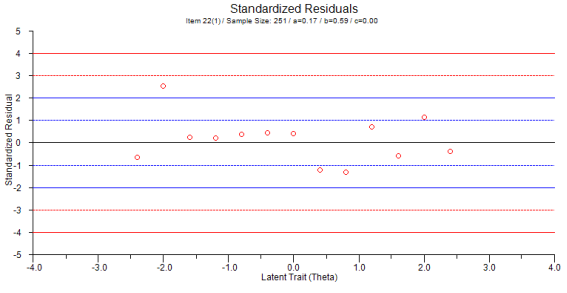

Standardized Residuals
Item 14(1) / Sample Size: 251 / a=0.23 / b=3.14 / c=0.00
Standardized Residual
Latent Trait (Theta)

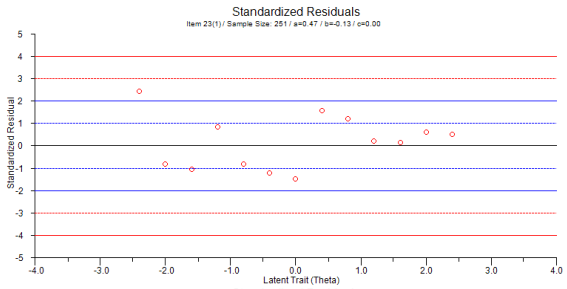

Standardized Residuals
Item 20(1) / Sample Size: 251 / a=0.11 / b=2.74 / c=0.00
Standardized Residual
Latent Trait (Theta)

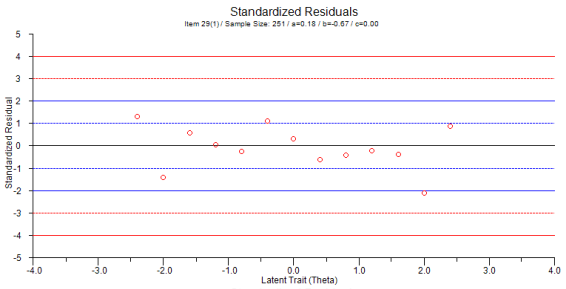

Standardized Residuals
Item 15(1) / Sample Size: 251 / a=0.43 / b=1.63 / c=0.00
Standardized Residual
Latent Trait (Theta)

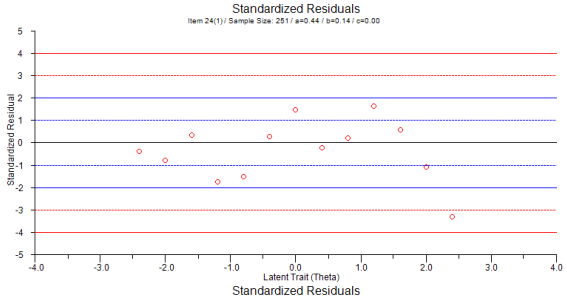

Standardized Residuals
Item 21(1) / Sample Size: 251 / a=0.16 / b=1.77 / c=0.00
Standardized Residual
Latent Trait (Theta)

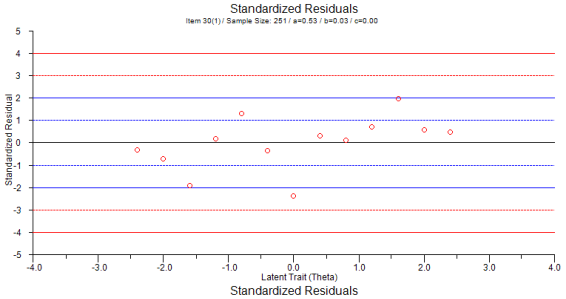

Standardized Residuals
Item 16(1) / Sample Size: 251 / a=0.42 / b=-0.01 / c=0.00
Standardized Residual
Latent Trait (Theta)

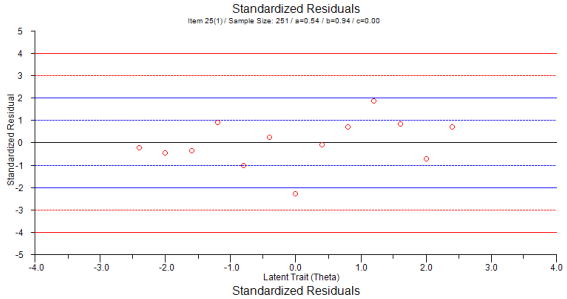

Standardized Residuals
Item 22(1) / Sample Size: 251 / a=0.17 / b=0.59 / c=0.00
Standardized Residual
Latent Trait (Theta)

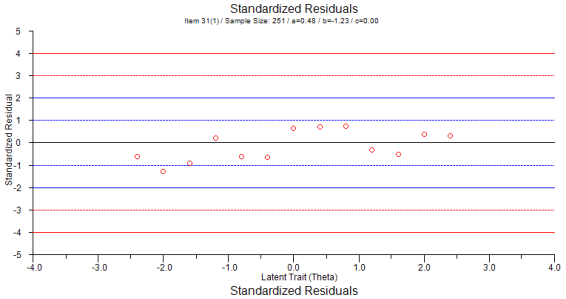

Standardized Residuals
Item 23(1) / Sample Size: 251 / a=0.47 / b=-0.13 / c=0.00
Standardized Residual
Latent Trait (Theta)

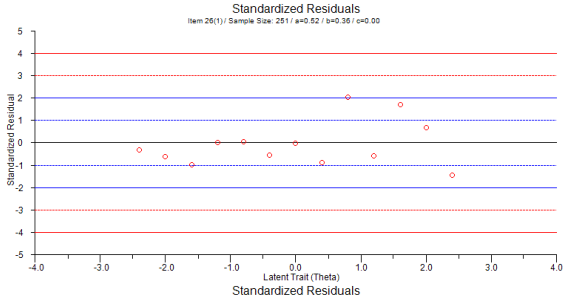

Standardized Residuals
Item 29(1) / Sample Size: 251 / a=0.18 / b=-0.67 / c=0.00
Standardized Residual
Latent Trait (Theta)

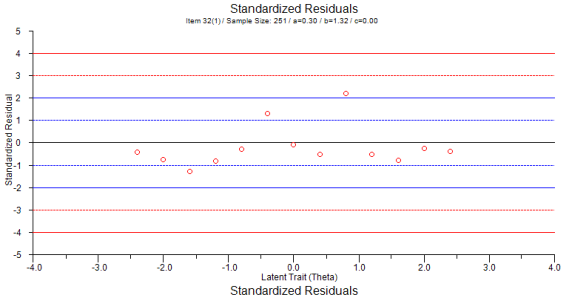

Standardized Residuals
Item 24(1) / Sample Size: 251 / a=0.44 / b=0.14 / c=0.00
Standardized Residual
Latent Trait (Theta)

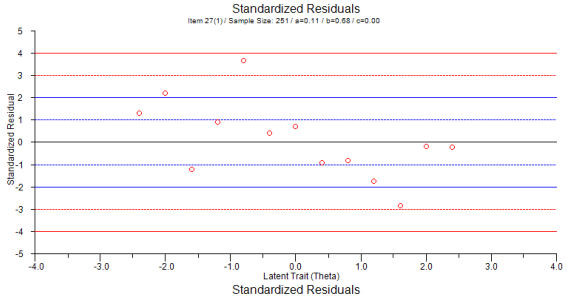

Standardized Residuals
Item 30(1) / Sample Size: 251 / a=0.53 / b=0.03 / c=0.00
Standardized Residual
Latent Trait (Theta)

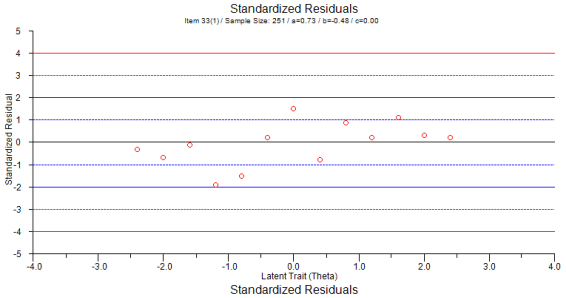

Standardized Residuals
Item 25(1) / Sample Size: 251 / a=0.54 / b=0.94 / c=0.00
Standardized Residual
Latent Trait (Theta)

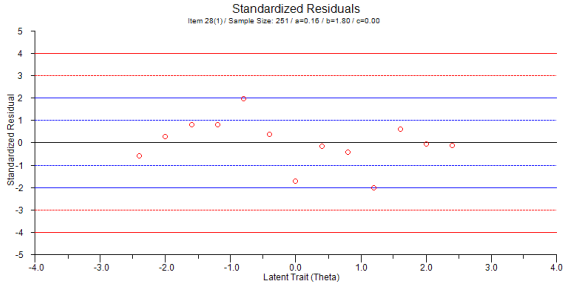

Standardized Residuals
Item 31(1) / Sample Size: 251 / a=0.48 / b=-1.23 / c=0.00
Standardized Residual
Latent Trait (Theta)

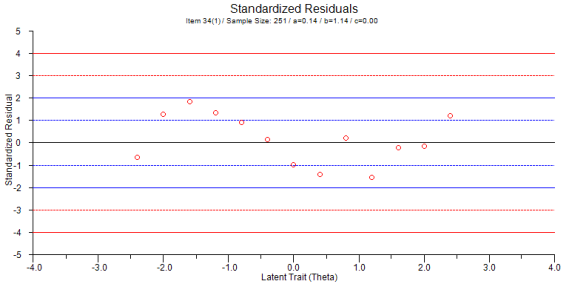

Standardized Residuals
Item 26(1) / Sample Size: 251 / a=0.52 / b=0.36 / c=0.00
Standardized Residual
Latent Trait (Theta)

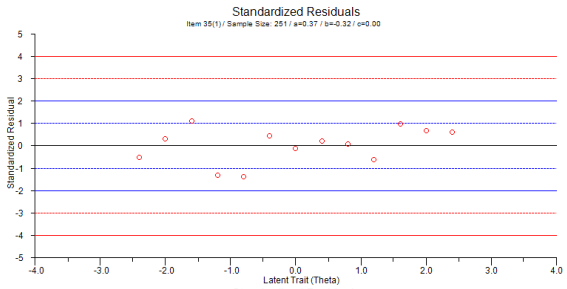

Standardized Residuals
Item 32(1) / Sample Size: 251 / a=0.30 / b=1.32 / c=0.00
Standardized Residual
Latent Trait (Theta)

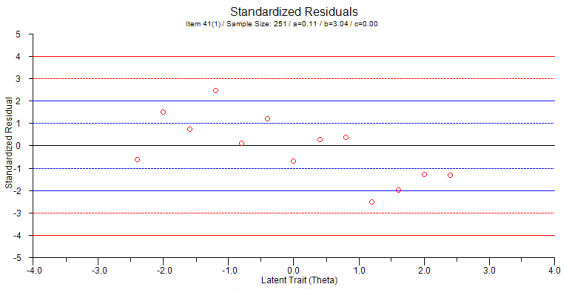

Standardized Residuals
Item 27(1) / Sample Size: 251 / a=0.11 / b=0.68 / c=0.00
Standardized Residual
Latent Trait (Theta)

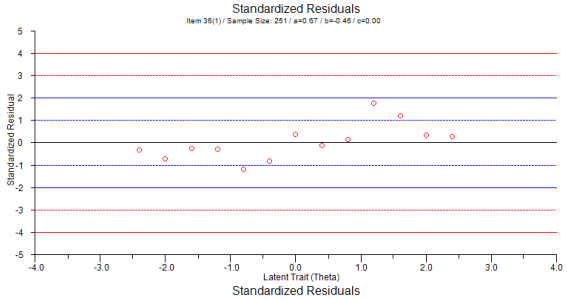

Standardized Residuals
Item 33(1) / Sample Size: 251 / a=0.73 / b=-0.48 / c=0.00
Standardized Residual
Latent Trait (Theta)

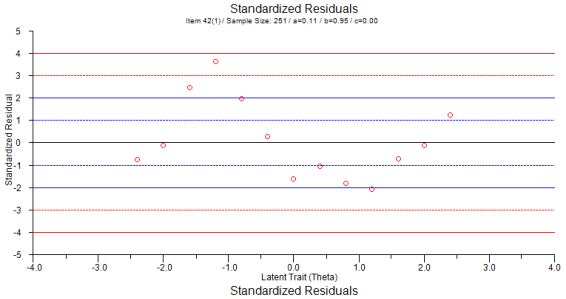

Standardized Residuals
Item 28(1) / Sample Size: 251 / a=0.16 / b=1.80 / c=0.00
Standardized Residual
Latent Trait (Theta)

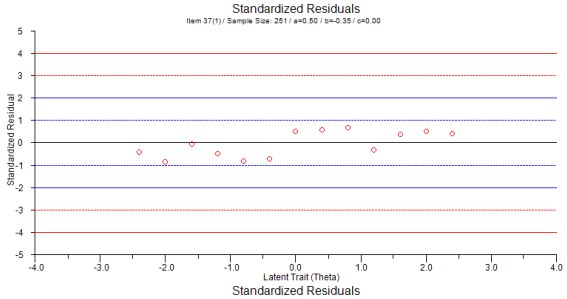

Standardized Residuals
Item 34(1) / Sample Size: 251 / a=0.14 / b=1.14 / c=0.00
Standardized Residual
Latent Trait (Theta)

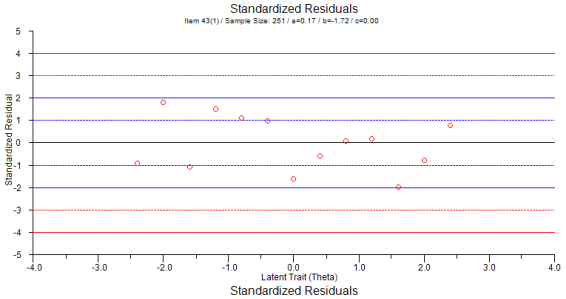

Standardized Residuals
Item 35(1) / Sample Size: 251 / a=0.37 / b=-0.32 / c=0.00
Standardized Residual
Latent Trait (Theta)

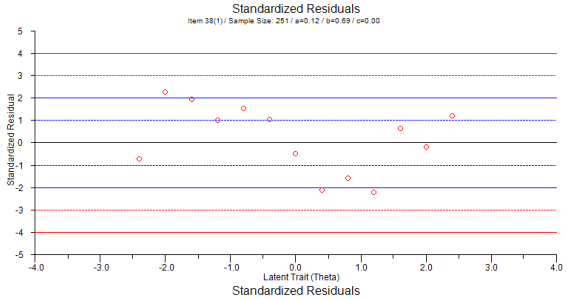

Standardized Residuals
Item 41(1) / Sample Size: 251 / a=0.11 / b=3.04 / c=0.00
Standardized Residual
Latent Trait (Theta)

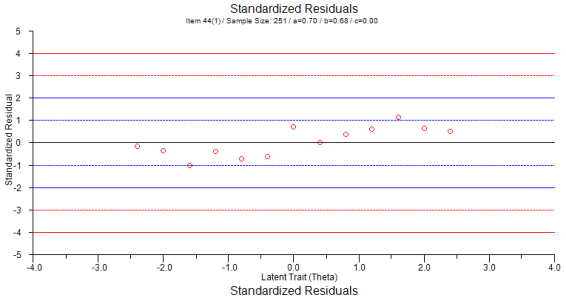

Standardized Residuals
Item 36(1) / Sample Size: 251 / a=0.67 / b=-0.46 / c=0.00
Standardized Residual
Latent Trait (Theta)

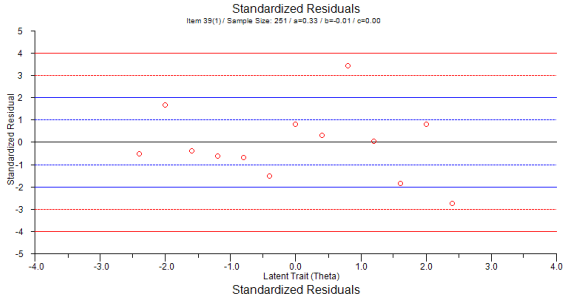

Standardized Residuals
Item 42(1) / Sample Size: 251 / a=0.11 / b=0.95 / c=0.00
Standardized Residual
Latent Trait (Theta)

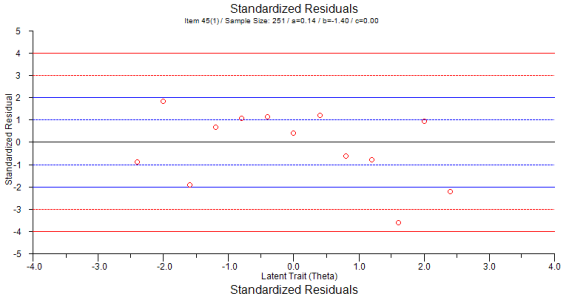

Standardized Residuals
Item 37(1) / Sample Size: 251 / a=0.50 / b=-0.35 / c=0.00
Standardized Residual
Latent Trait (Theta)

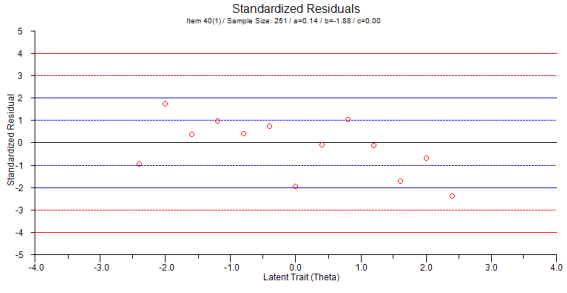

Standardized Residuals
Item 43(1) / Sample Size: 251 / a=0.17 / b=-1.72 / c=0.00
Standardized Residual
Latent Trait (Theta)

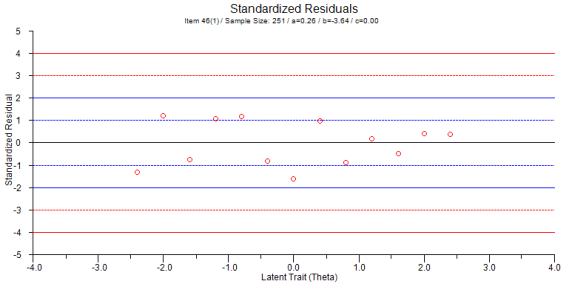

Standardized Residuals
Item 38(1) / Sample Size: 251 / a=0.12 / b=0.69 / c=0.00
Standardized Residual
Latent Trait (Theta)

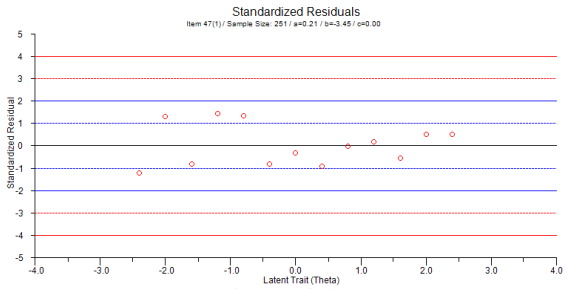

Standardized Residuals
Item 44(1) / Sample Size: 251 / a=0.70 / b=0.68 / c=0.00
Standardized Residual
Latent Trait (Theta)

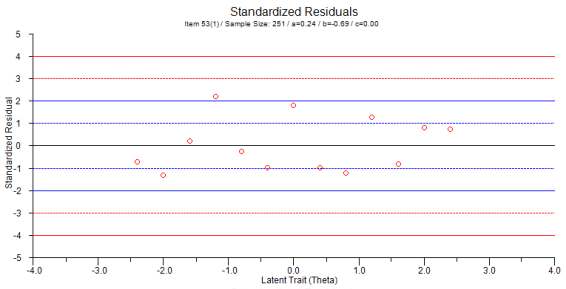

Standardized Residuals
Item 39(1) / Sample Size: 251 / a=0.33 / b=-0.01 / c=0.00
Standardized Residual
Latent Trait (Theta)

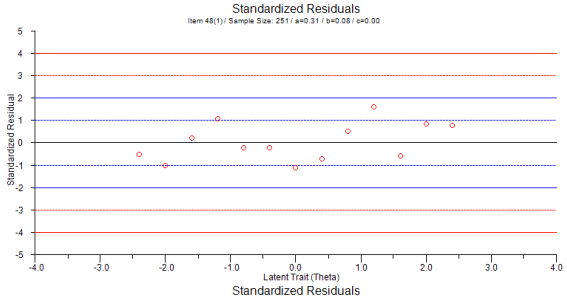

Standardized Residuals
Item 45(1) / Sample Size: 251 / a=0.14 / b=-1.40 / c=0.00
Standardized Residual
Latent Trait (Theta)

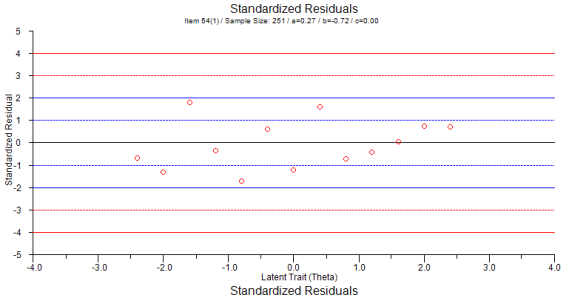

Standardized Residuals
Item 40(1) / Sample Size: 251 / a=0.14 / b=-1.88 / c=0.00
Standardized Residual
Latent Trait (Theta)

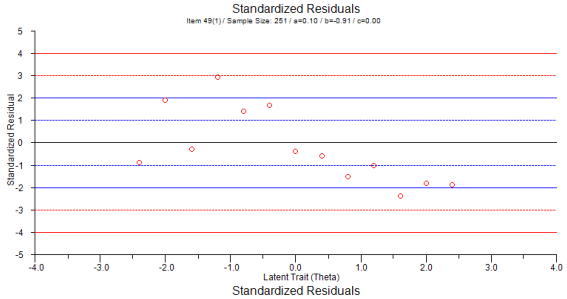

Standardized Residuals
Item 46(1) / Sample Size: 251 / a=0.26 / b=-3.64 / c=0.00
Standardized Residual
Latent Trait (Theta)

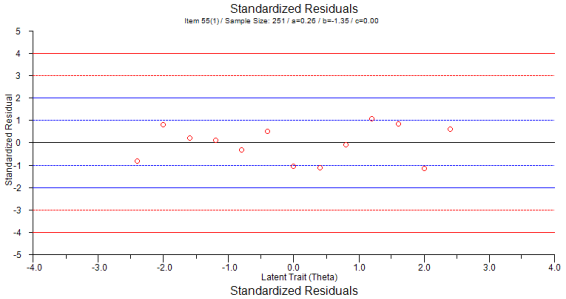

Standardized Residuals
Item 47(1) / Sample Size: 251 / a=0.21 / b=-3.45 / c=0.00
Standardized Residual
Latent Trait (Theta)

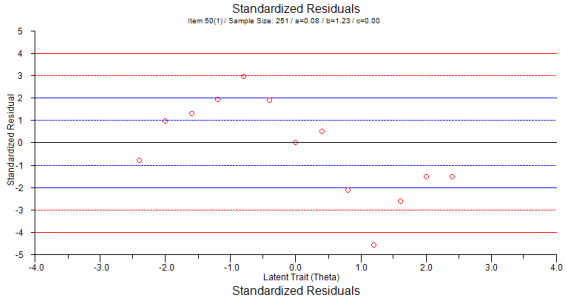

Standardized Residuals
Item 53(1) / Sample Size: 251 / a=0.24 / b=-0.69 / c=0.00
Standardized Residual
Latent Trait (Theta)

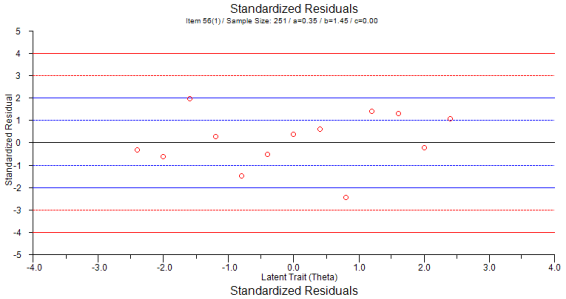

Standardized Residuals
Item 48(1) / Sample Size: 251 / a=0.31 / b=0.08 / c=0.00
Standardized Residual
Latent Trait (Theta)

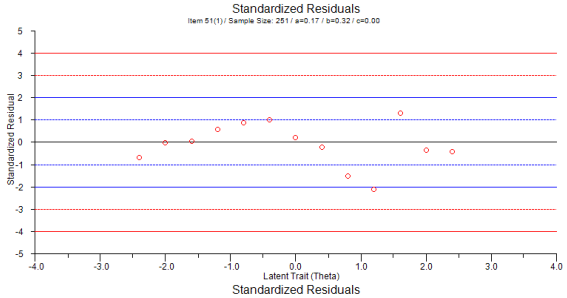

Standardized Residuals
Item 54(1) / Sample Size: 251 / a=0.27 / b=-0.72 / c=0.00
Standardized Residual
Latent Trait (Theta)

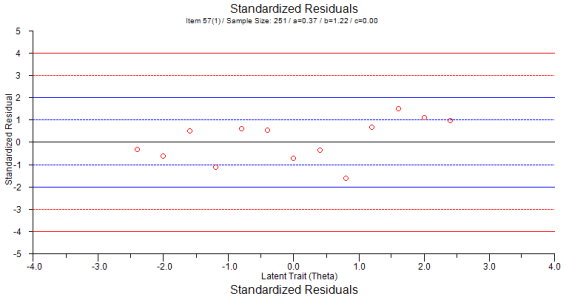

Standardized Residuals
Item 49(1) / Sample Size: 251 / a=0.10 / b=-0.91 / c=0.00
Standardized Residual
Latent Trait (Theta)

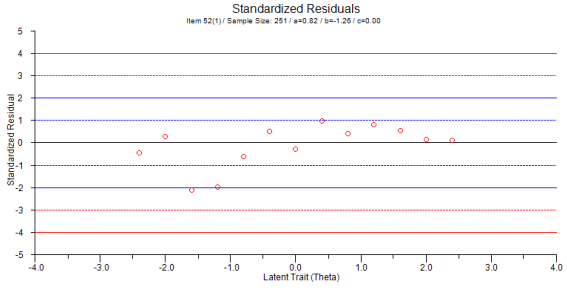

Standardized Residuals
Item 55(1) / Sample Size: 251 / a=0.26 / b=-1.35 / c=0.00
Standardized Residual
Latent Trait (Theta)

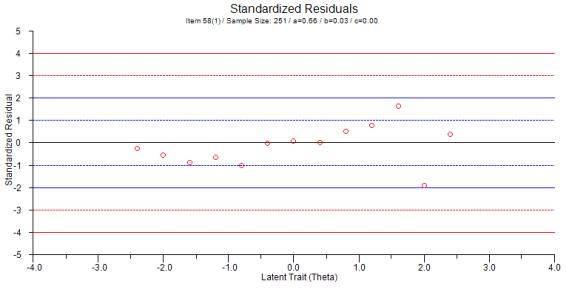

Standardized Residuals
Item 50(1) / Sample Size: 251 / a=0.08 / b=1.23 / c=0.00
Standardized Residual
Latent Trait (Theta)

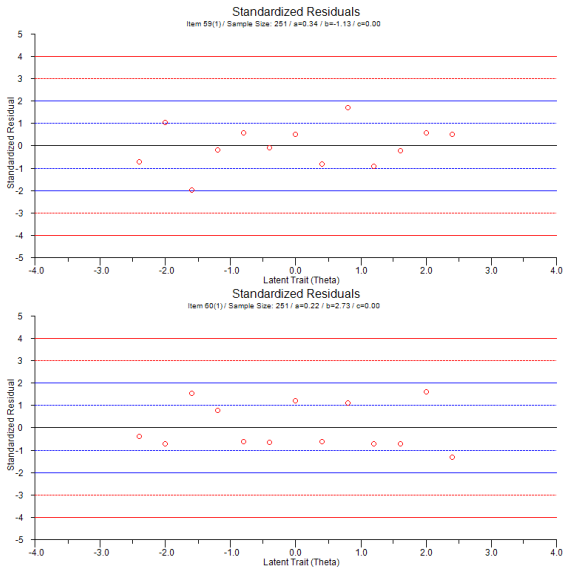

Standardized Residuals
Item 56(1) / Sample Size: 251 / a=0.35 / b=1.45 / c=0.00
Standardized Residual
Latent Trait (Theta)


Standardized Residuals
Item 51(1) / Sample Size: 251 / a=0.17 / b=0.32 / c=0.00
Standardized Residual
Latent Trait (Theta)


Standardized Residuals
Item 57(1) / Sample Size: 251 / a=0.37 / b=1.22 / c=0.00
Standardized Residual
Latent Trait (Theta)

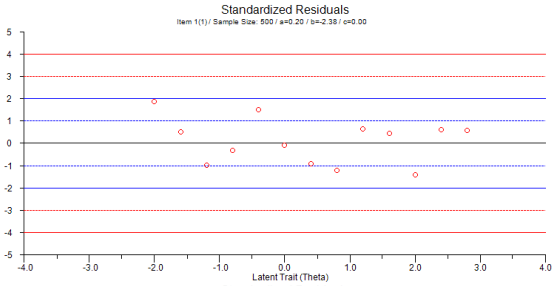

Standardized Residuals
Item 52(1) / Sample Size: 251 / a=0.82 / b=-1.26 / c=0.00
Standardized Residual
Latent Trait (Theta)

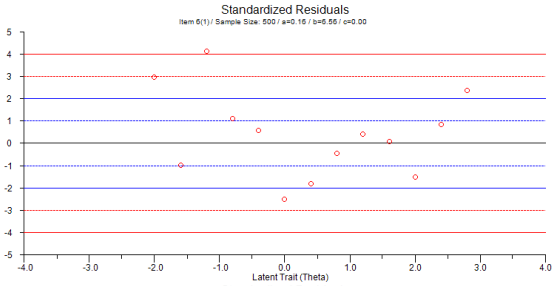

Standardized Residuals
Item 58(1) / Sample Size: 251 / a=0.66 / b=0.03 / c=0.00
Standardized Residual
Latent Trait (Theta)

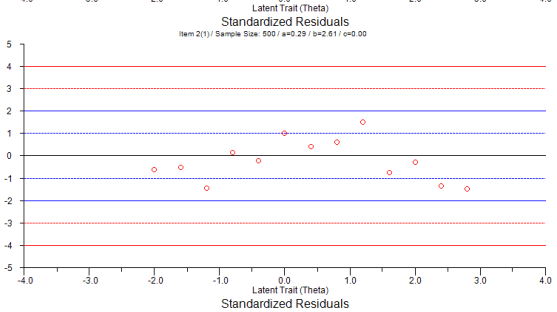

Standardized Residuals
Item 59(1) / Sample Size: 251 / a=0.34 / b=-1.13 / c=0.00
Standardized Residual
Latent Trait (Theta)

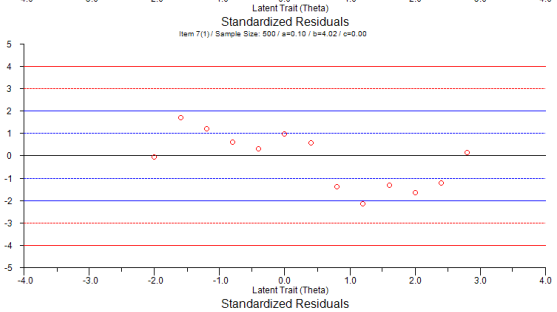
Standardized Residuals
Item 60(1) / Sample Size: 251 / a=0.22 / b=2.73 / c=0.00
Standardized Residual
Latent Trait (Theta)

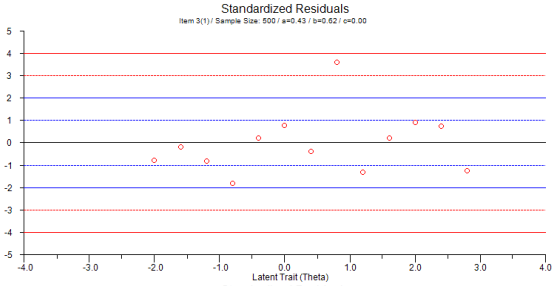
Appendix D: SRs 2plm sample 500

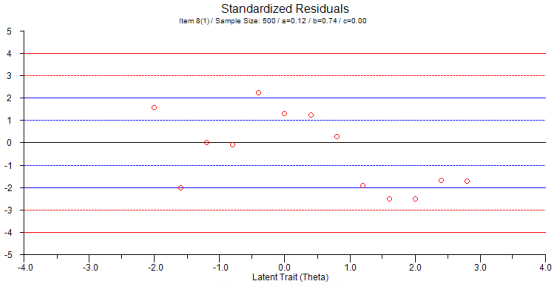


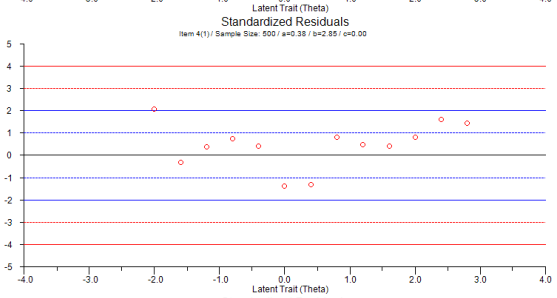


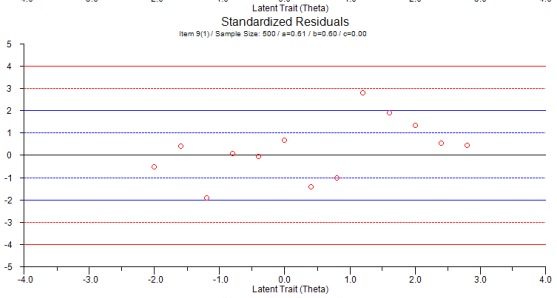


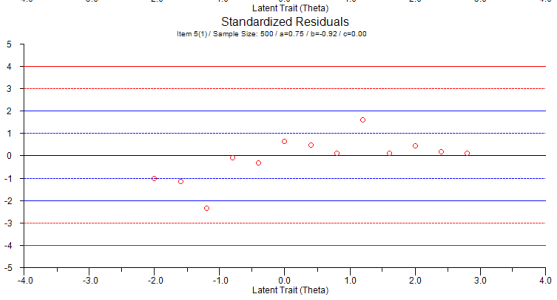


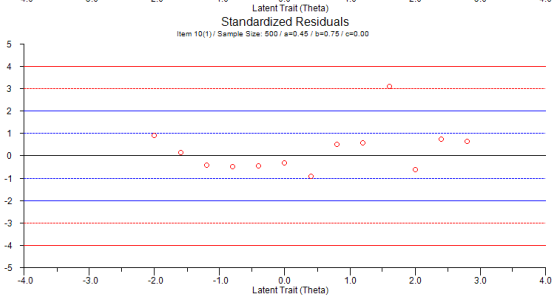


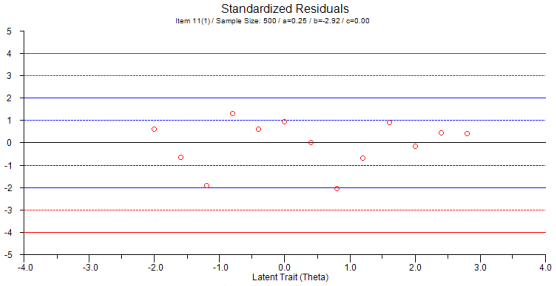


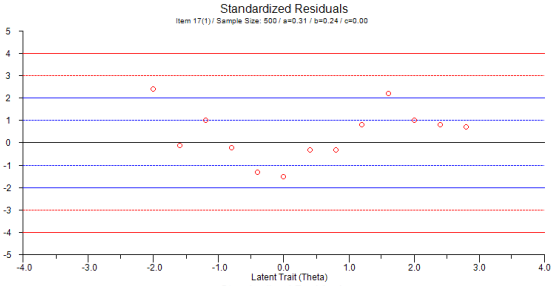


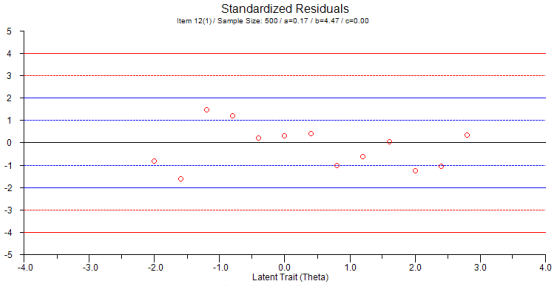


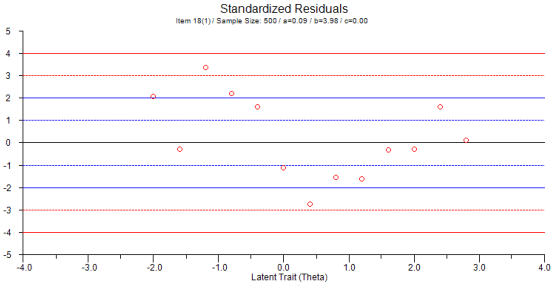


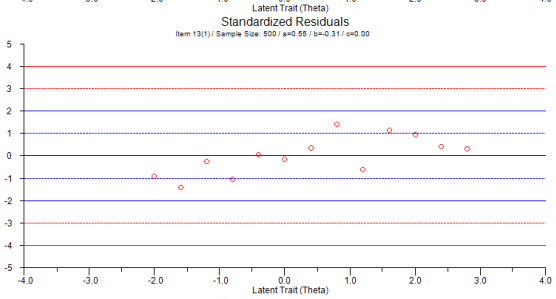


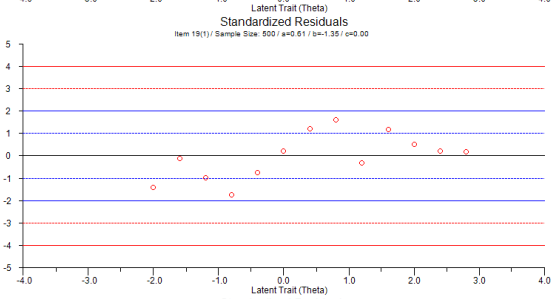

Standardized Residuals
Item 11(1) / Sample Size: 500 / a=0.25 / b=-2.92 / c=0.00
Latent Trait (Theta)

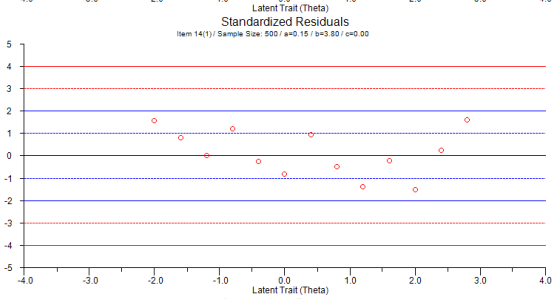

Standardized Residuals
Item 17(1) / Sample Size: 500 / a=0.31 / b=0.24 / c=0.00
Latent Trait (Theta)

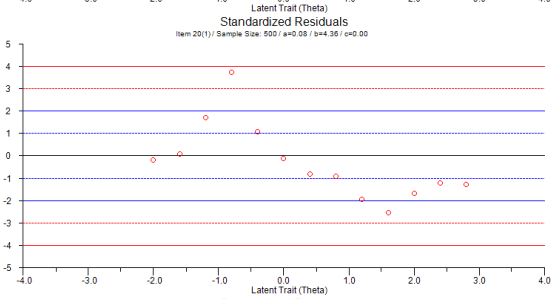

Standardized Residuals
Item 12(1) / Sample Size: 500 / a=0.17 / b=4.47 / c=0.00
Latent Trait (Theta)

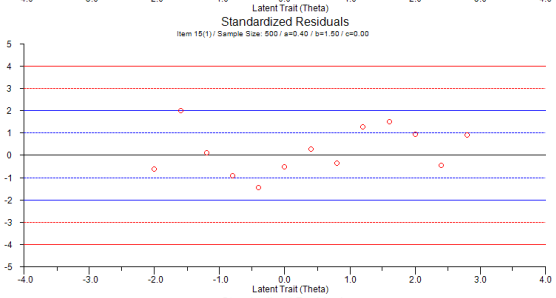

Standardized Residuals
Item 18(1) / Sample Size: 500 / a=0.09 / b=3.98 / c=0.00
Latent Trait (Theta)

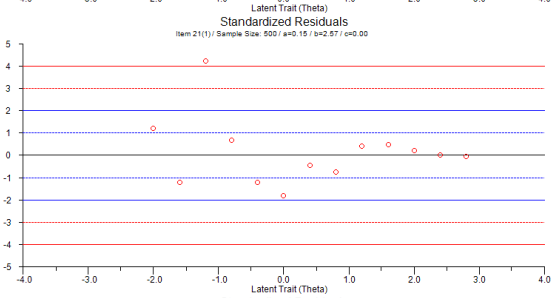

Standardized Residuals
Item 13(1) / Sample Size: 500 / a=0.55 / b=-0.31 / c=0.00
Latent Trait (Theta)

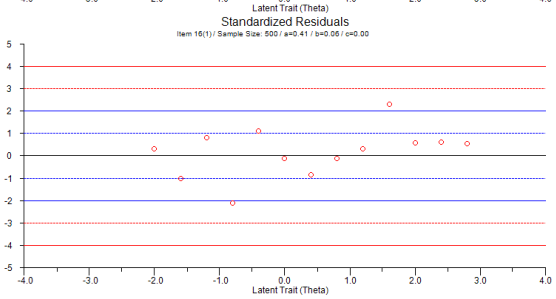

Standardized Residuals
Item 19(1) / Sample Size: 500 / a=0.61 / b=-1.35 / c=0.00
Latent Trait (Theta)

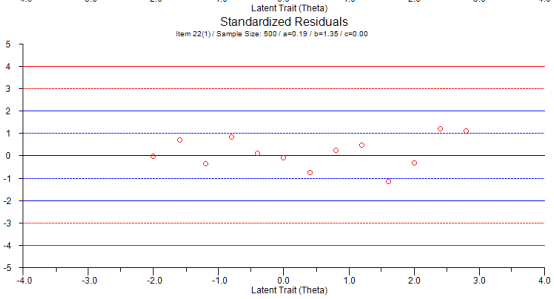

Standardized Residuals
Item 14(1) / Sample Size: 500 / a=0.15 / b=3.80 / c=0.00
Latent Trait (Theta)

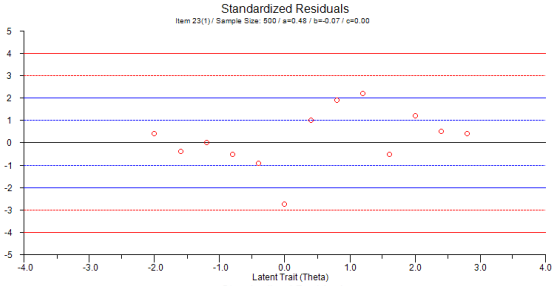

Standardized Residuals
Item 20(1) / Sample Size: 500 / a=0.08 / b=4.36 / c=0.00
Latent Trait (Theta)

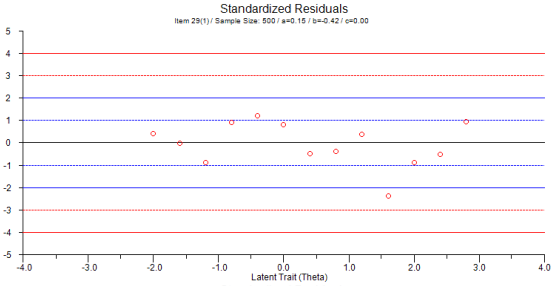

Standardized Residuals
Item 15(1) / Sample Size: 500 / a=0.40 / b=1.50 / c=0.00
Latent Trait (Theta)

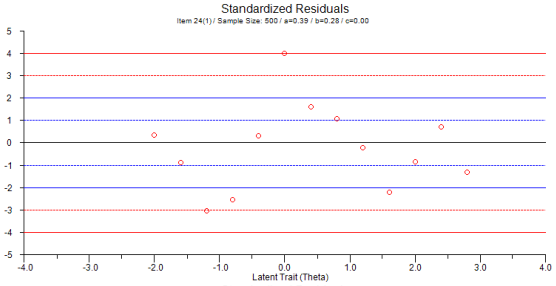

Standardized Residuals
Item 21(1) / Sample Size: 500 / a=0.15 / b=2.57 / c=0.00
Latent Trait (Theta)

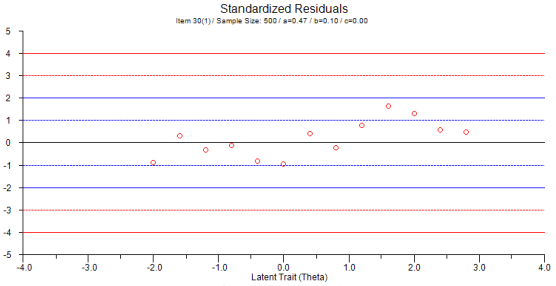

Standardized Residuals
Item 16(1) / Sample Size: 500 / a=0.41 / b=0.06 / c=0.00
Latent Trait (Theta)

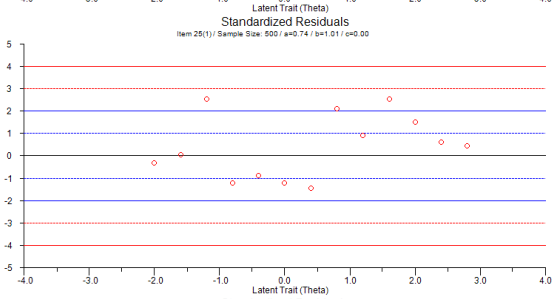

Standardized Residuals
Item 22(1) / Sample Size: 500 / a=0.19 / b=1.35 / c=0.00
Latent Trait (Theta)

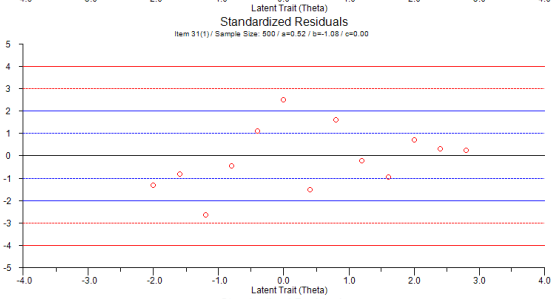

Standardized Residuals
Item 23(1) / Sample Size: 500 / a=0.48 / b=-0.07 / c=0.00
Latent Trait (Theta)

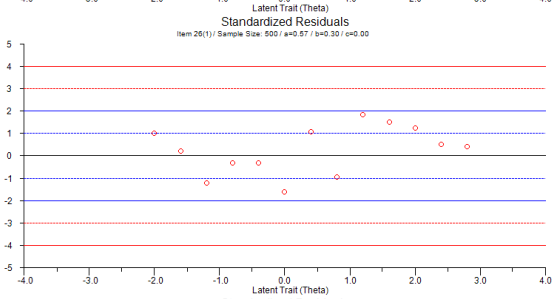

Standardized Residuals
Item 29(1) / Sample Size: 500 / a=0.15 / b=-0.42 / c=0.00
Latent Trait (Theta)

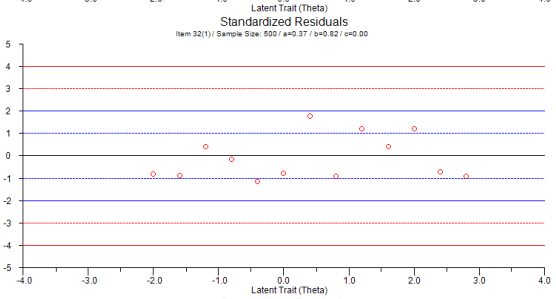

Standardized Residuals
Item 24(1) / Sample Size: 500 / a=0.39 / b=0.28 / c=0.00
Latent Trait (Theta)

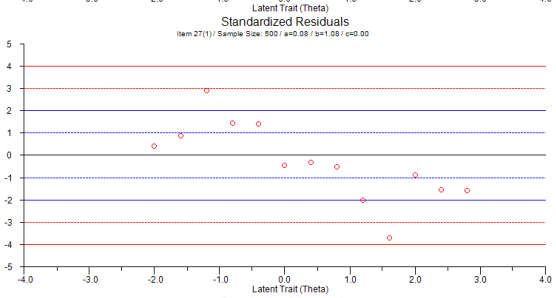

Standardized Residuals
Item 30(1) / Sample Size: 500 / a=0.47 / b=0.10 / c=0.00
Latent Trait (Theta)

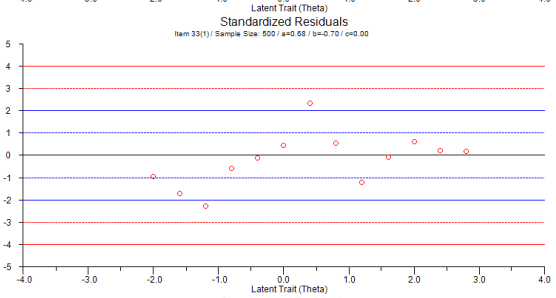

Standardized Residuals
Item 25(1) / Sample Size: 500 / a=0.74 / b=1.01 / c=0.00
Latent Trait (Theta)

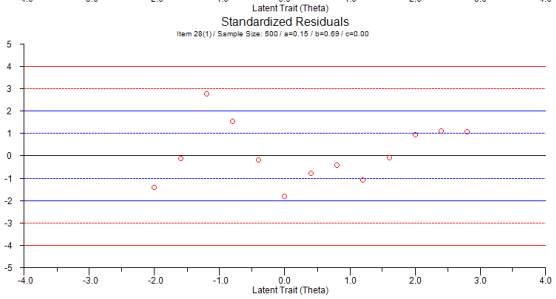

Standardized Residuals
Item 31(1) / Sample Size: 500 / a=0.52 / b=-1.08 / c=0.00
Latent Trait (Theta)

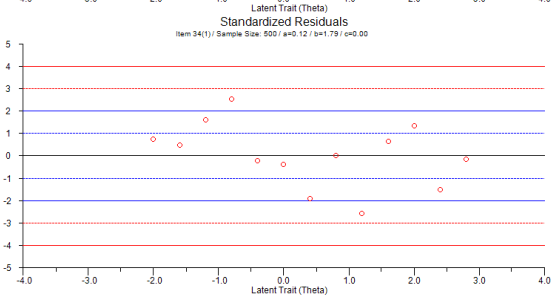

Standardized Residuals
Item 26(1) / Sample Size: 500 / a=0.57 / b=0.30 / c=0.00
Latent Trait (Theta)

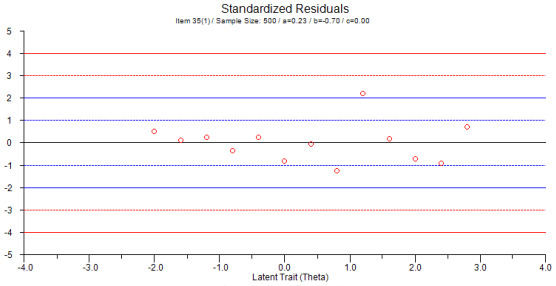

Standardized Residuals
Item 32(1) / Sample Size: 500 / a=0.37 / b=0.82 / c=0.00
Latent Trait (Theta)

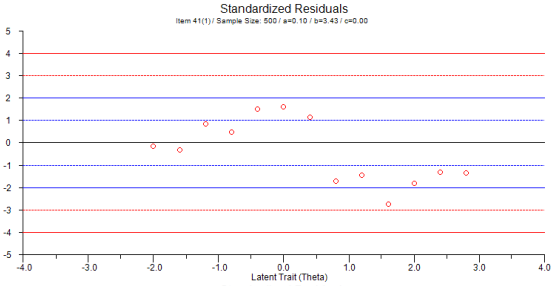

Standardized Residuals
Item 27(1) / Sample Size: 500 / a=0.08 / b=1.08 / c=0.00
Latent Trait (Theta)

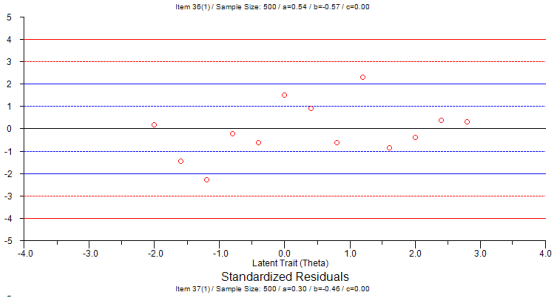

Standardized Residuals
Item 33(1) / Sample Size: 500 / a=0.68 / b=-0.70 / c=0.00
Latent Trait (Theta)

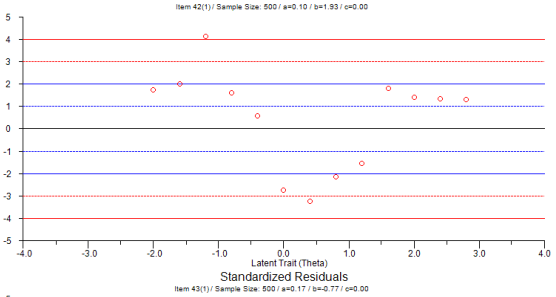

Standardized Residuals
Item 28(1) / Sample Size: 500 / a=0.15 / b=0.69 / c=0.00
Latent Trait (Theta)

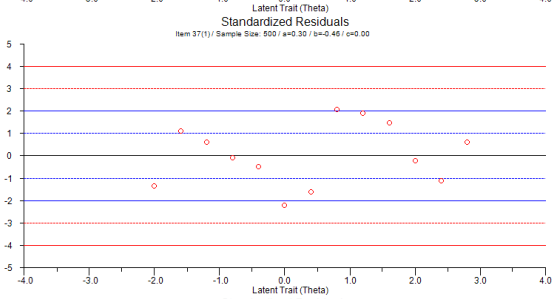

Standardized Residuals
Item 34(1) / Sample Size: 500 / a=0.12 / b=1.79 / c=0.00
Latent Trait (Theta)

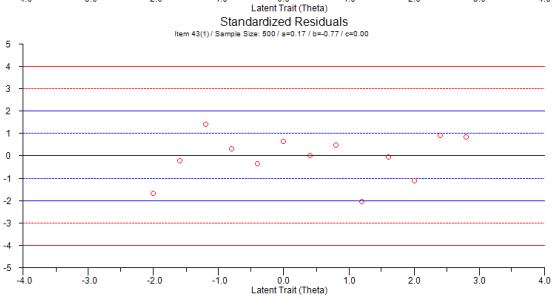

Standardized Residuals
Item 35(1) / Sample Size: 500 / a=0.23 / b=-0.70 / c=0.00
Latent Trait (Theta)

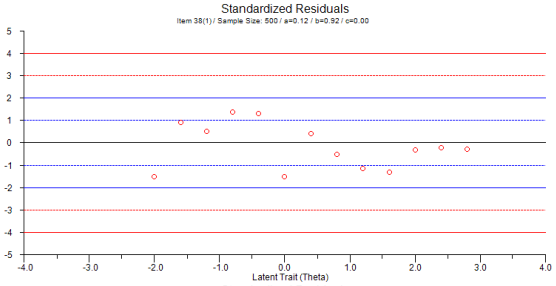

Standardized Residuals
Item 41(1) / Sample Size: 500 / a=0.10 / b=3.43 / c=0.00
Latent Trait (Theta)

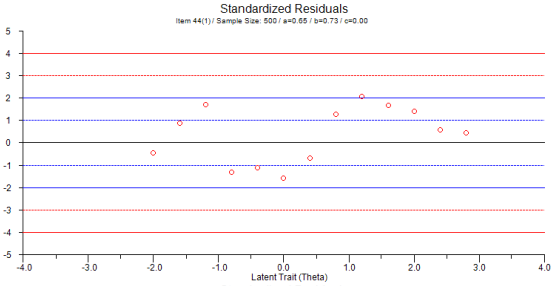

Standardized Residuals
Item 36(1) / Sample Size: 500 / a=0.54 / b=-0.57 / c=0.00
Latent Trait (Theta)

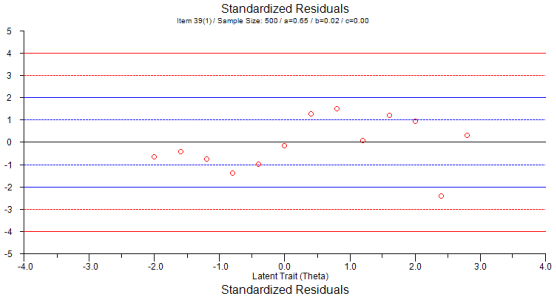

Standardized Residuals
Item 42(1) / Sample Size: 500 / a=0.10 / b=1.93 / c=0.00
Latent Trait (Theta)

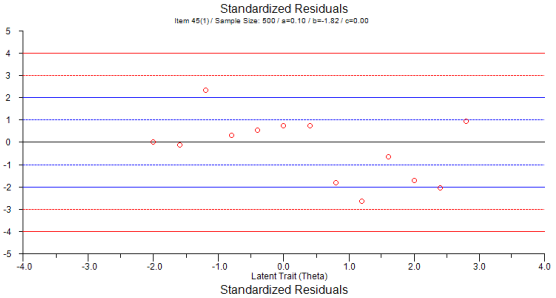

Standardized Residuals
Item 37(1) / Sample Size: 500 / a=0.30 / b=-0.46 / c=0.00
Latent Trait (Theta)

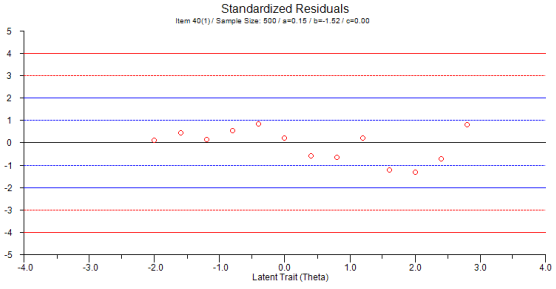

Standardized Residuals
Item 43(1) / Sample Size: 500 / a=0.17 / b=-0.77 / c=0.00
Latent Trait (Theta)

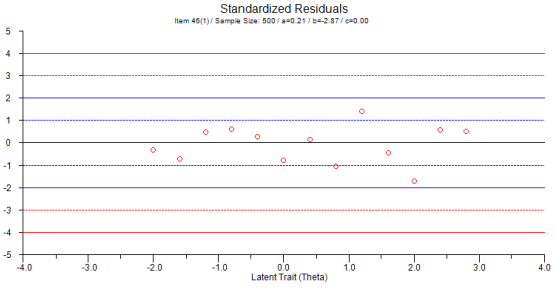

Standardized Residuals
Item 38(1) / Sample Size: 500 / a=0.12 / b=0.92 / c=0.00
Latent Trait (Theta)

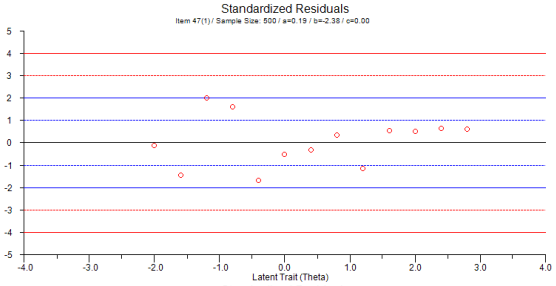

Standardized Residuals
Item 44(1) / Sample Size: 500 / a=0.65 / b=0.73 / c=0.00
Latent Trait (Theta)

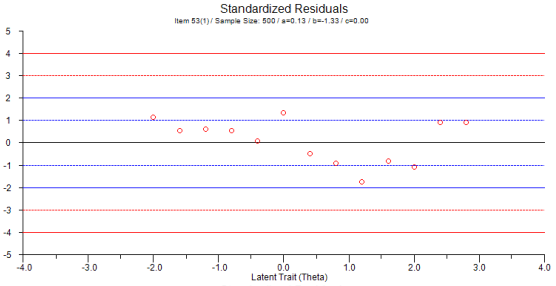

Standardized Residuals
Item 39(1) / Sample Size: 500 / a=0.65 / b=0.02 / c=0.00
Latent Trait (Theta)

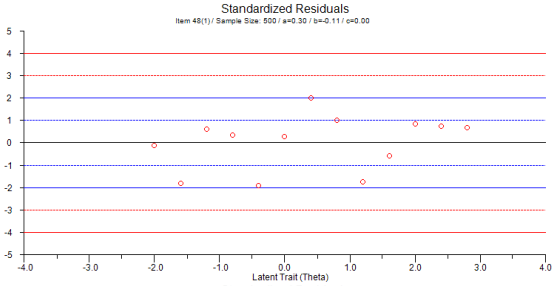

Standardized Residuals
Item 45(1) / Sample Size: 500 / a=0.10 / b=-1.82 / c=0.00
Latent Trait (Theta)

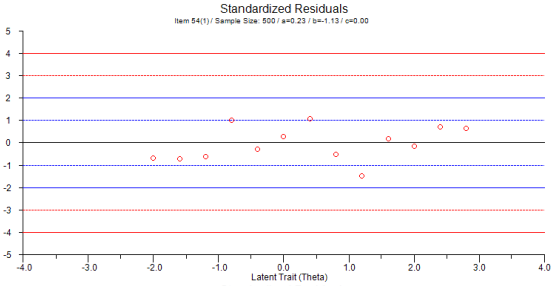

Standardized Residuals
Item 40(1) / Sample Size: 500 / a=0.15 / b=-1.52 / c=0.00
Latent Trait (Theta)

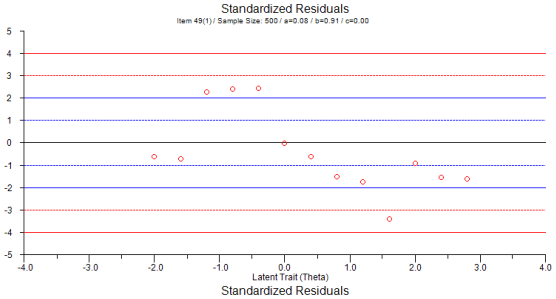

Standardized Residuals
Item 46(1) / Sample Size: 500 / a=0.21 / b=-2.87 / c=0.00
Latent Trait (Theta)

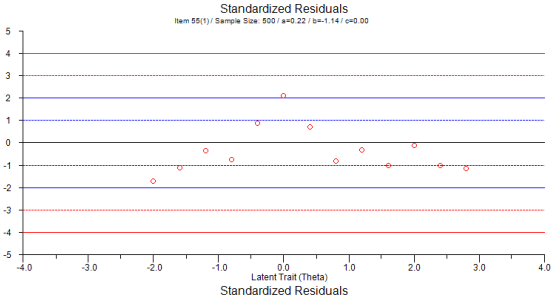

Standardized Residuals
Item 47(1) / Sample Size: 500 / a=0.19 / b=-2.38 / c=0.00
Latent Trait (Theta)

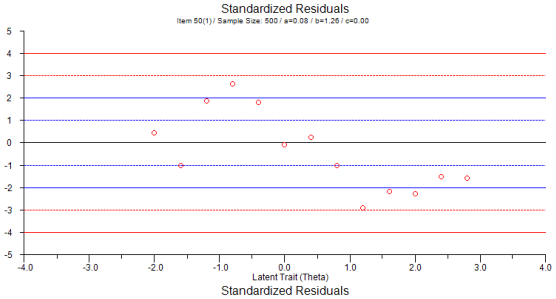

Standardized Residuals
Item 53(1) / Sample Size: 500 / a=0.13 / b=-1.33 / c=0.00
Latent Trait (Theta)

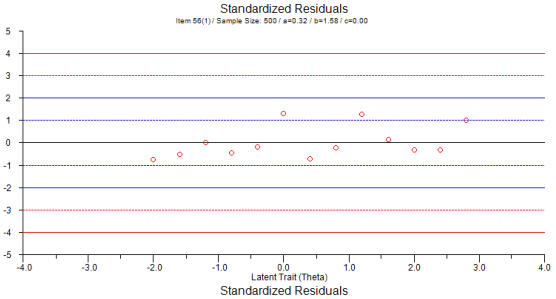

Standardized Residuals
Item 48(1) / Sample Size: 500 / a=0.30 / b=-0.11 / c=0.00
Latent Trait (Theta)

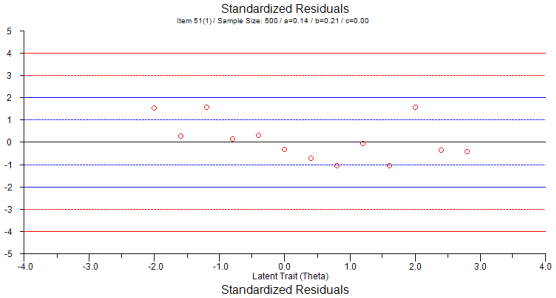

Standardized Residuals
Item 54(1) / Sample Size: 500 / a=0.23 / b=-1.13 / c=0.00
Latent Trait (Theta)

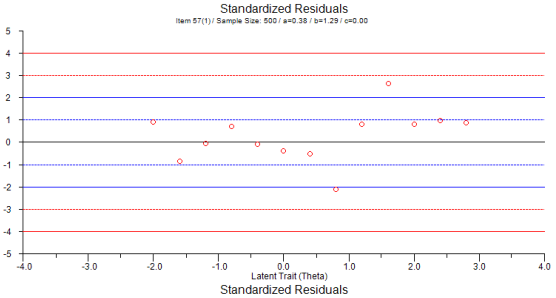

Standardized Residuals
Item 49(1) / Sample Size: 500 / a=0.08 / b=0.91 / c=0.00
Latent Trait (Theta)

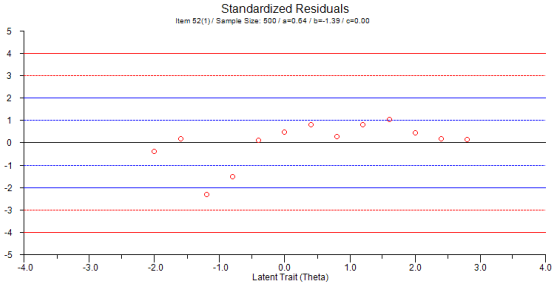

Standardized Residuals
Item 55(1) / Sample Size: 500 / a=0.22 / b=-1.14 / c=0.00
Latent Trait (Theta)

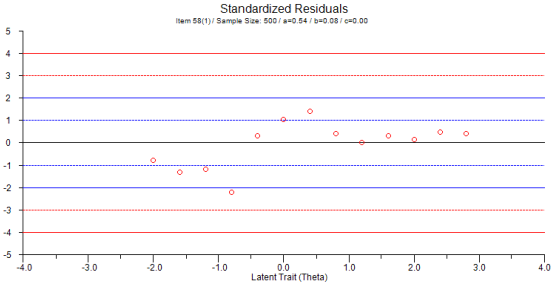

Standardized Residuals
Item 50(1) / Sample Size: 500 / a=0.08 / b=1.26 / c=0.00
Latent Trait (Theta)

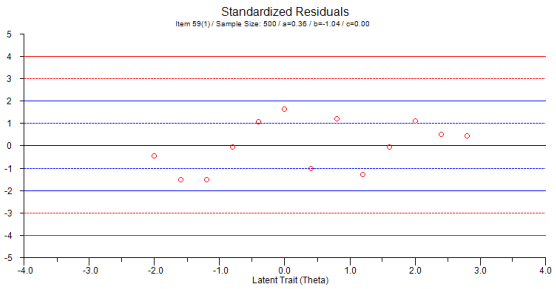

Standardized Residuals
Item 56(1) / Sample Size: 500 / a=0.32 / b=1.58 / c=0.00
Latent Trait (Theta)

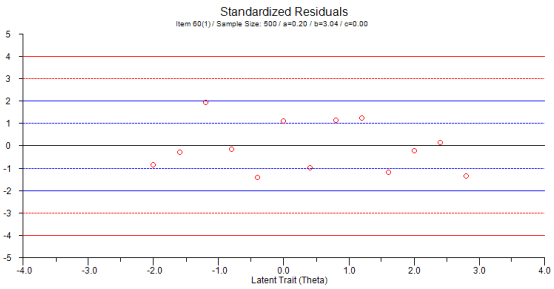

Standardized Residuals
Item 51(1) / Sample Size: 500 / a=0.14 / b=0.21 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 57(1) / Sample Size: 500 / a=0.38 / b=1.29 / c=0.00
Latent Trait (Theta)

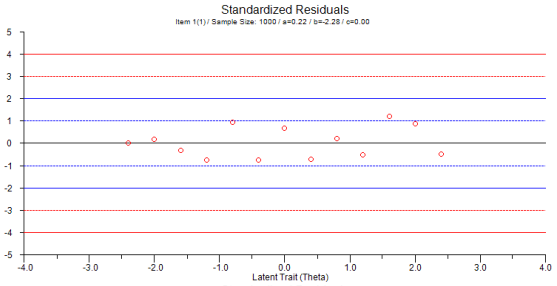

Standardized Residuals
Item 52(1) / Sample Size: 500 / a=0.64 / b=-1.39 / c=0.00
Latent Trait (Theta)

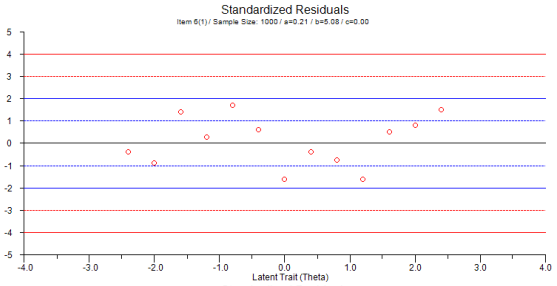

Standardized Residuals
Item 58(1) / Sample Size: 500 / a=0.54 / b=0.08 / c=0.00
Latent Trait (Theta)

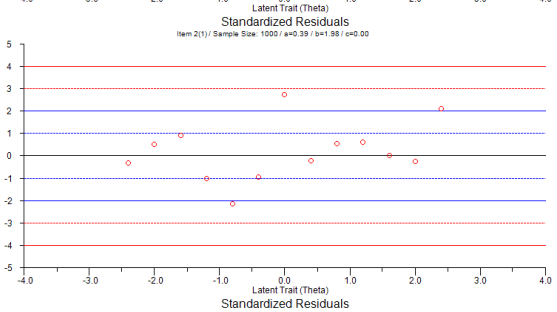

Standardized Residuals
Item 59(1) / Sample Size: 500 / a=0.36 / b=-1.04 / c=0.00
5
4
3
2
1
0
-1
-2
-3
-4
-5
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

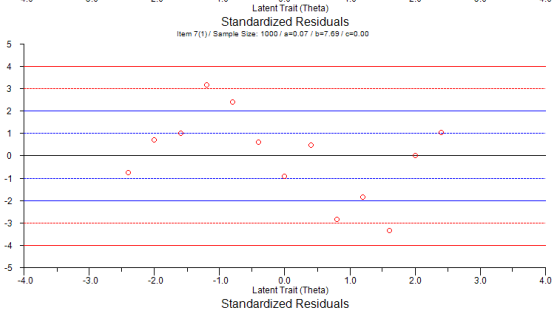

Standardized Residuals
Item 60(1) / Sample Size: 500 / a=0.20 / b=3.04 / c=0.00
5
4
3
2
1
0
-1
-2
-3
-4
-5
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent Trait (Theta)

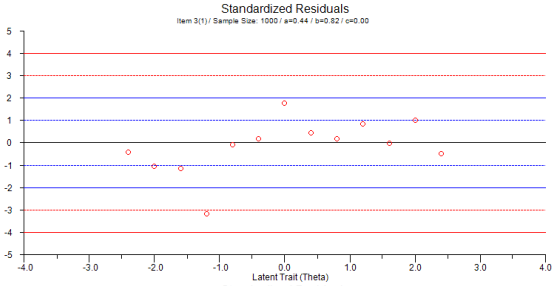
Appendix D: SRs 2plm for sample 1000

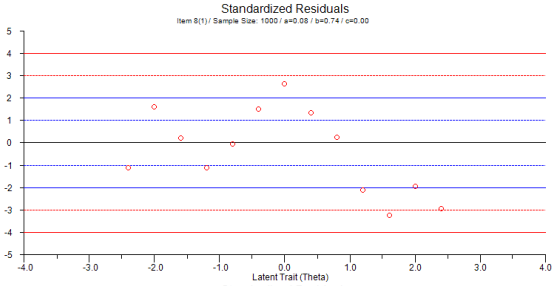


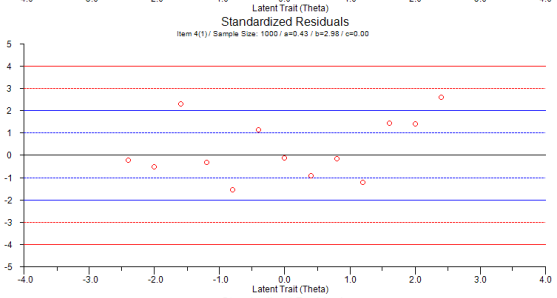


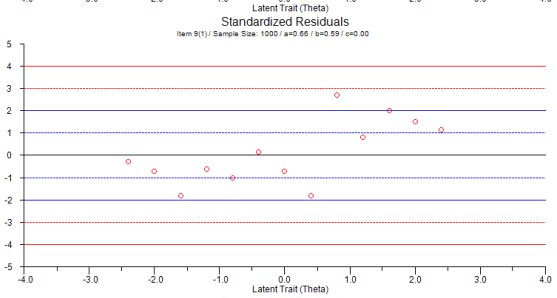


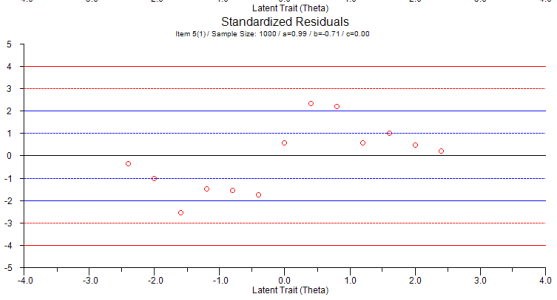


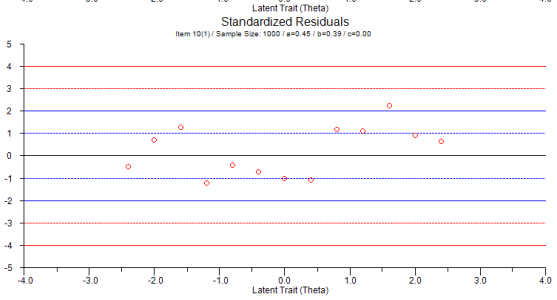


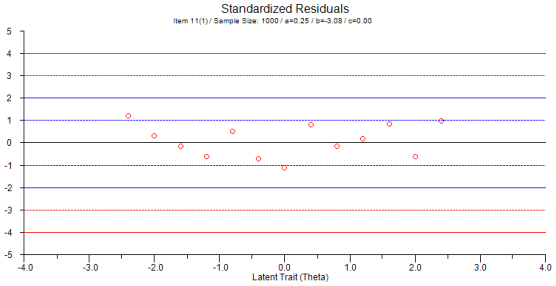


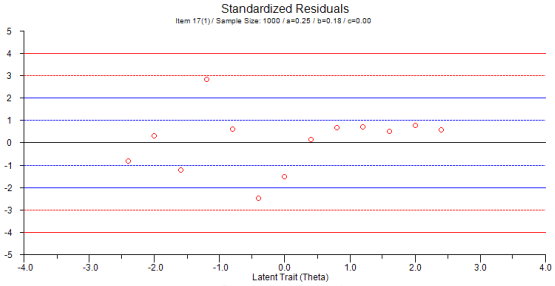


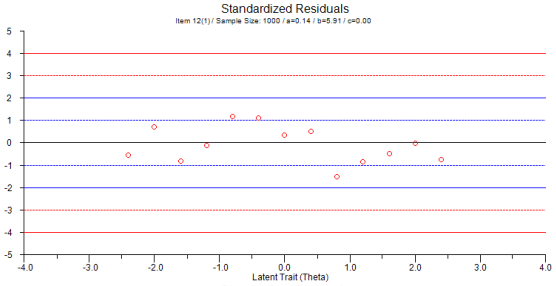


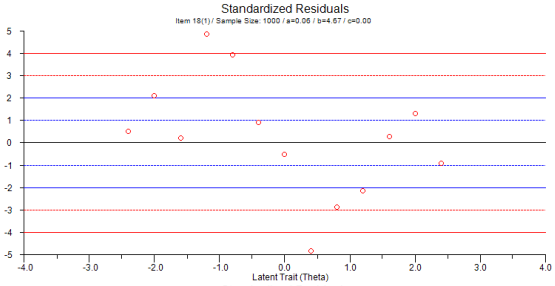


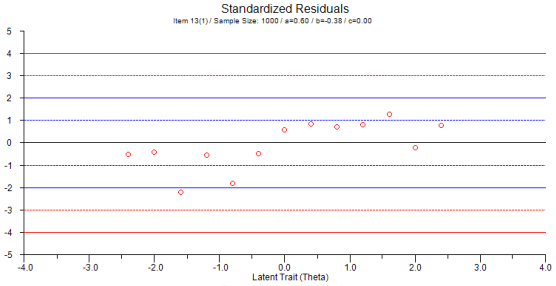


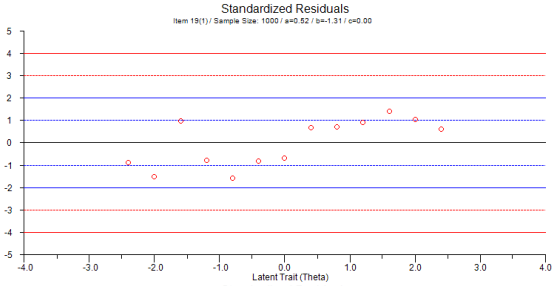

Standardized Residuals
Item 11(1) / Sample Size: 1000 / a=0.25 / b=-3.08 / c=0.00
Latent Trait (Theta)

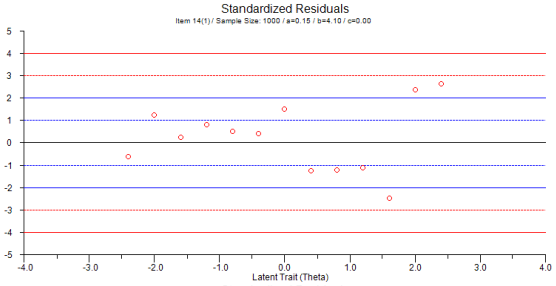

Standardized Residuals
Item 17(1) / Sample Size: 1000 / a=0.25 / b=0.18 / c=0.00
Latent Trait (Theta)

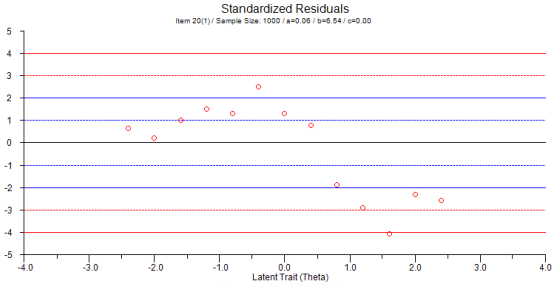

Standardized Residuals
Item 12(1) / Sample Size: 1000 / a=0.14 / b=5.91 / c=0.00
Latent Trait (Theta)

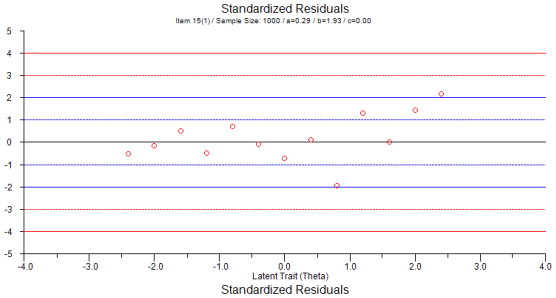

Standardized Residuals
Item 18(1) / Sample Size: 1000 / a=0.06 / b=4.67 / c=0.00
Latent Trait (Theta)

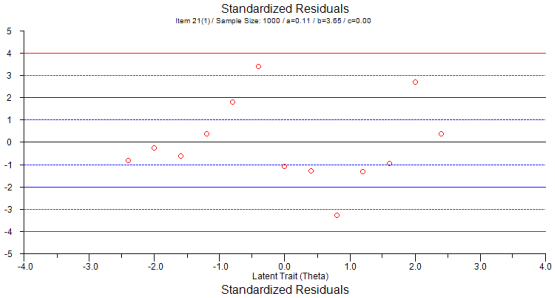

Standardized Residuals
Item 13(1) / Sample Size: 1000 / a=0.60 / b=-0.38 / c=0.00
Latent Trait (Theta)

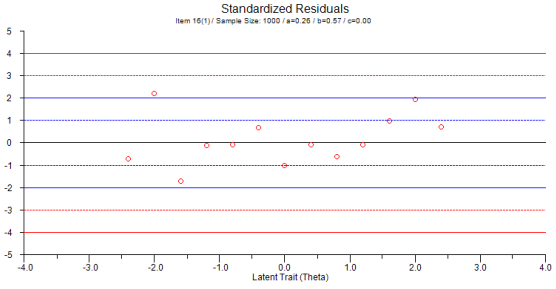

Standardized Residuals
Item 19(1) / Sample Size: 1000 / a=0.52 / b=-1.31 / c=0.00
Latent Trait (Theta)

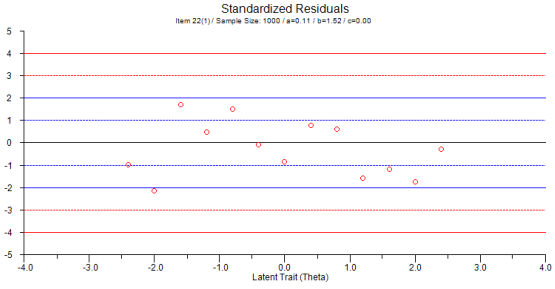

Standardized Residuals
Item 14(1) / Sample Size: 1000 / a=0.15 / b=4.10 / c=0.00
Latent Trait (Theta)

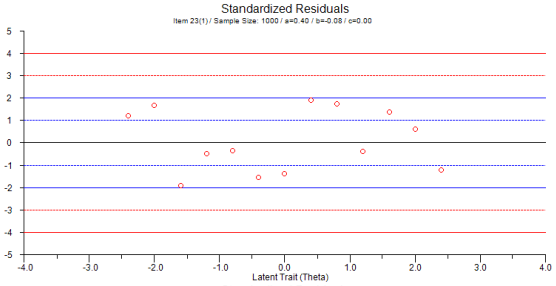

Standardized Residuals
Item 20(1) / Sample Size: 1000 / a=0.06 / b=6.54 / c=0.00
Latent Trait (Theta)

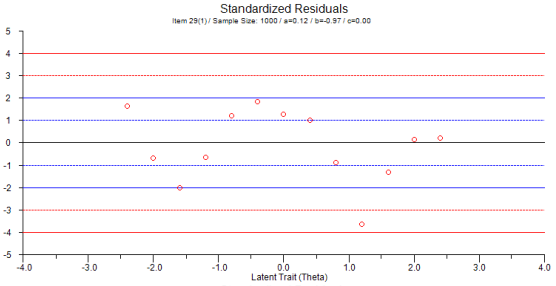

Standardized Residuals
Item 15(1) / Sample Size: 1000 / a=0.29 / b=1.93 / c=0.00
Latent Trait (Theta)

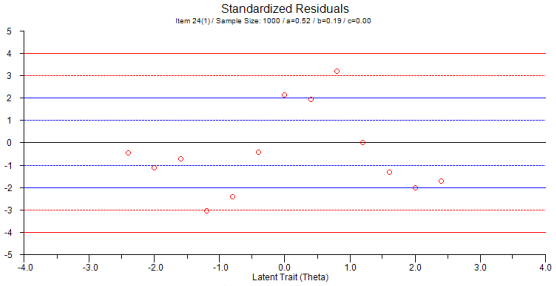

Standardized Residuals
Item 21(1) / Sample Size: 1000 / a=0.11 / b=3.65 / c=0.00
Latent Trait (Theta)

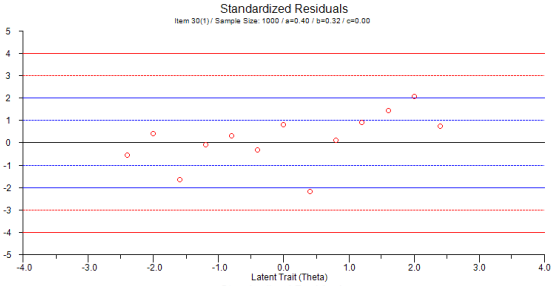

Standardized Residuals
Item 16(1) / Sample Size: 1000 / a=0.26 / b=0.57 / c=0.00
Latent Trait (Theta)

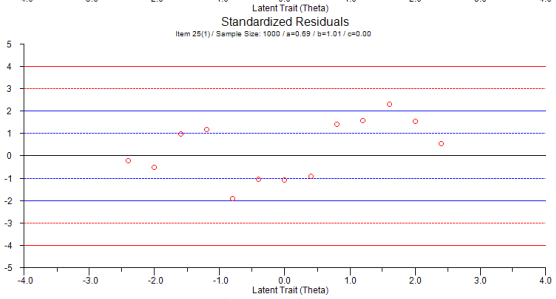

Standardized Residuals
Item 22(1) / Sample Size: 1000 / a=0.11 / b=1.52 / c=0.00
Latent Trait (Theta)

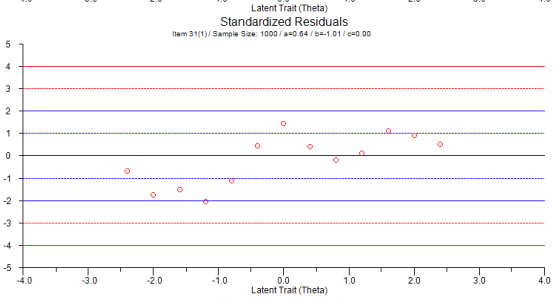

Standardized Residuals
Item 23(1) / Sample Size: 1000 / a=0.40 / b=-0.08 / c=0.00
Latent Trait (Theta)

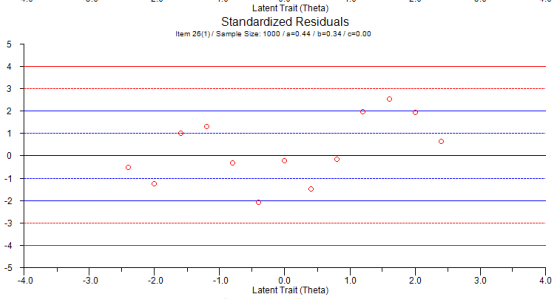

Standardized Residuals
Item 29(1) / Sample Size: 1000 / a=0.12 / b=-0.97 / c=0.00
Latent Trait (Theta)

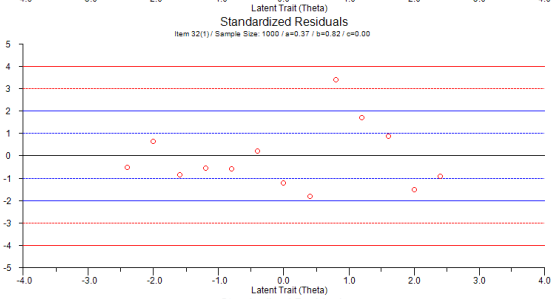

Standardized Residuals
Item 24(1) / Sample Size: 1000 / a=0.52 / b=0.19 / c=0.00
Latent Trait (Theta)

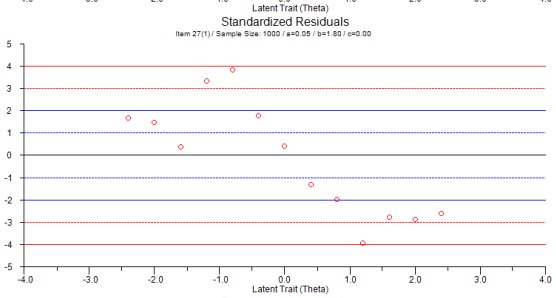

Standardized Residuals
Item 30(1) / Sample Size: 1000 / a=0.40 / b=0.32 / c=0.00
Latent Trait (Theta)

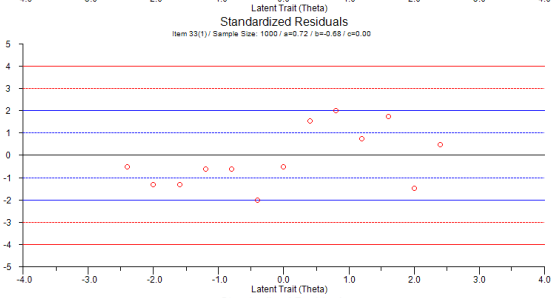

Standardized Residuals
Item 25(1) / Sample Size: 1000 / a=0.69 / b=1.01 / c=0.00
Latent Trait (Theta)

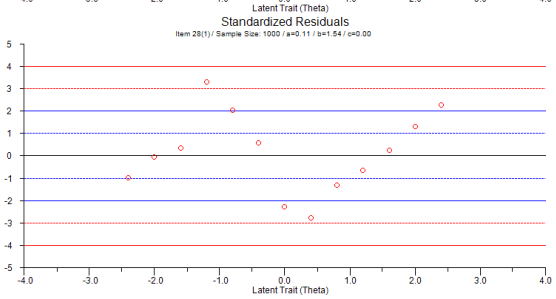

Standardized Residuals
Item 31(1) / Sample Size: 1000 / a=0.64 / b=-1.01 / c=0.00
Latent Trait (Theta)

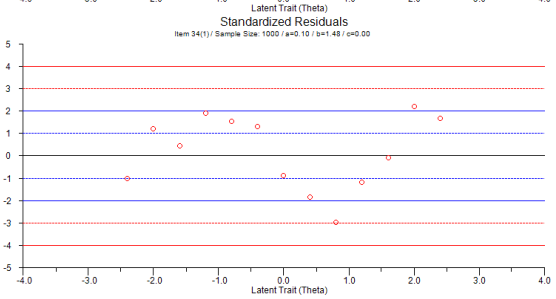

Standardized Residuals
Item 26(1) / Sample Size: 1000 / a=0.44 / b=0.34 / c=0.00
Latent Trait (Theta)

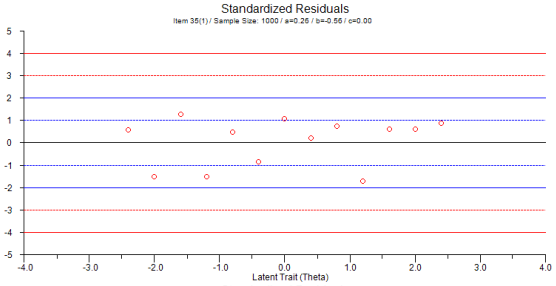

Standardized Residuals
Item 32(1) / Sample Size: 1000 / a=0.37 / b=0.82 / c=0.00
Latent Trait (Theta)

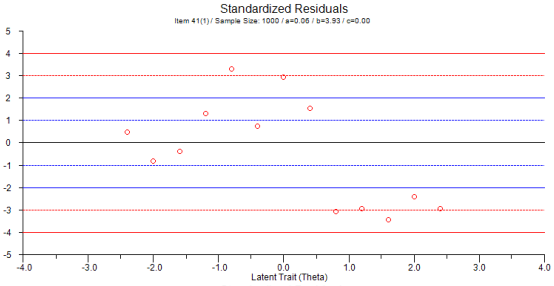

Standardized Residuals
Item 27(1) / Sample Size: 1000 / a=0.05 / b=1.80 / c=0.00
Latent Trait (Theta)

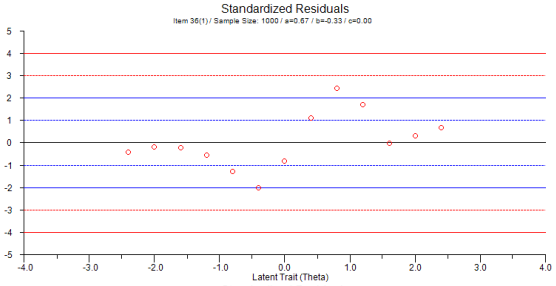

Standardized Residuals
Item 33(1) / Sample Size: 1000 / a=0.72 / b=-0.68 / c=0.00
Latent Trait (Theta)

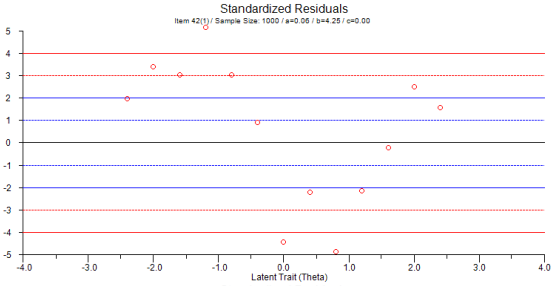

Standardized Residuals
Item 28(1) / Sample Size: 1000 / a=0.11 / b=1.54 / c=0.00
Latent Trait (Theta)

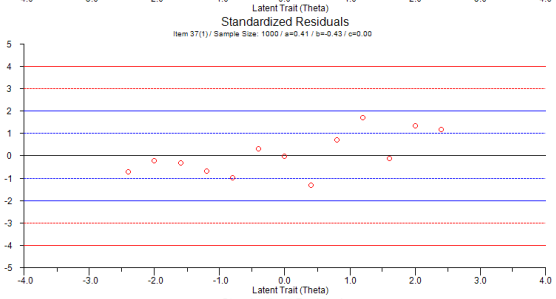

Standardized Residuals
Item 34(1) / Sample Size: 1000 / a=0.10 / b=1.48 / c=0.00
Latent Trait (Theta)

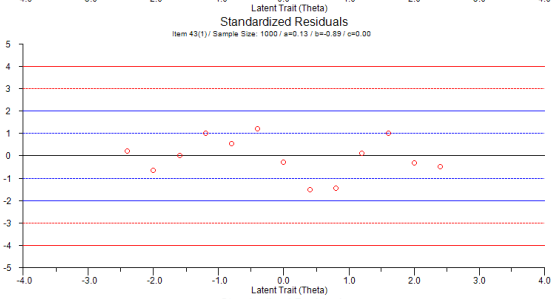

Standardized Residuals
Item 35(1) / Sample Size: 1000 / a=0.26 / b=-0.56 / c=0.00
Latent Trait (Theta)

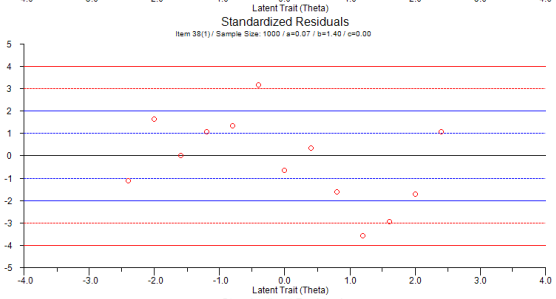

Standardized Residuals
Item 41(1) / Sample Size: 1000 / a=0.06 / b=3.93 / c=0.00
Latent Trait (Theta)

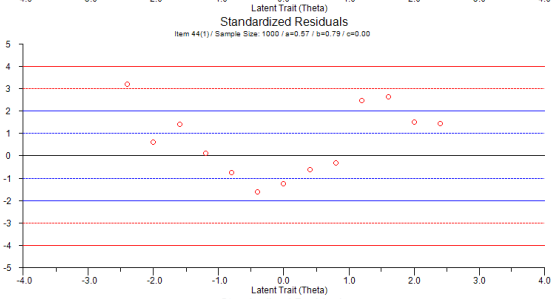

Standardized Residuals
Item 36(1) / Sample Size: 1000 / a=0.67 / b=-0.33 / c=0.00
Latent Trait (Theta)

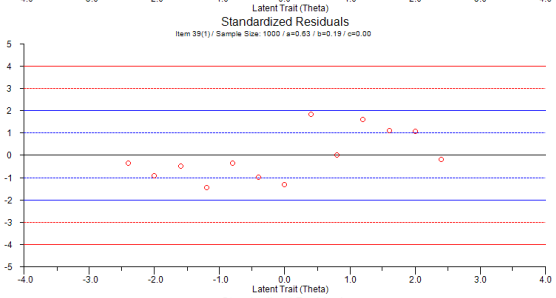

Standardized Residuals
Item 42(1) / Sample Size: 1000 / a=0.06 / b=4.25 / c=0.00
Latent Trait (Theta)

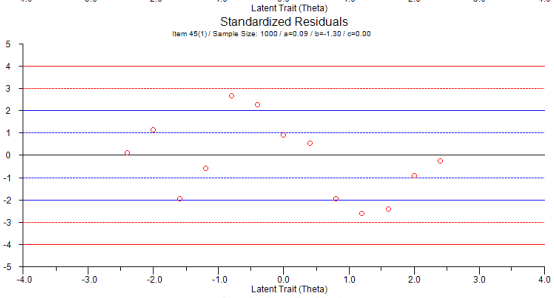

Standardized Residuals
Item 37(1) / Sample Size: 1000 / a=0.41 / b=-0.43 / c=0.00
Latent Trait (Theta)

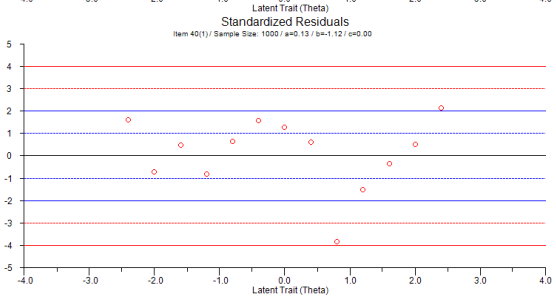

Standardized Residuals
Item 43(1) / Sample Size: 1000 / a=0.13 / b=-0.89 / c=0.00
Latent Trait (Theta)

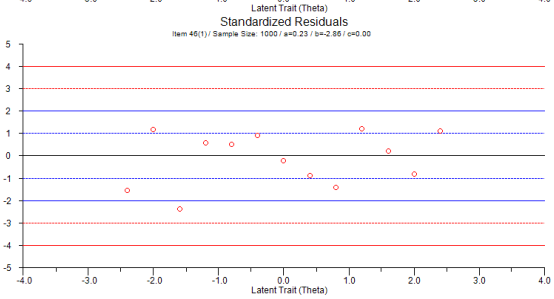

Standardized Residuals
Item 38(1) / Sample Size: 1000 / a=0.07 / b=1.40 / c=0.00
Latent Trait (Theta)

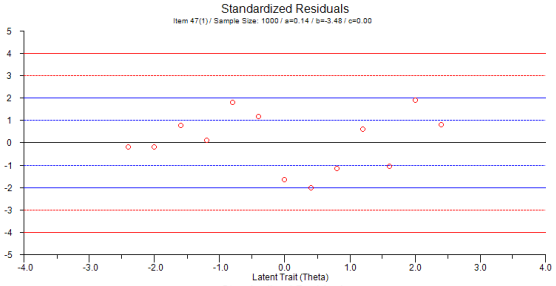

Standardized Residuals
Item 44(1) / Sample Size: 1000 / a=0.57 / b=0.79 / c=0.00
Latent Trait (Theta)

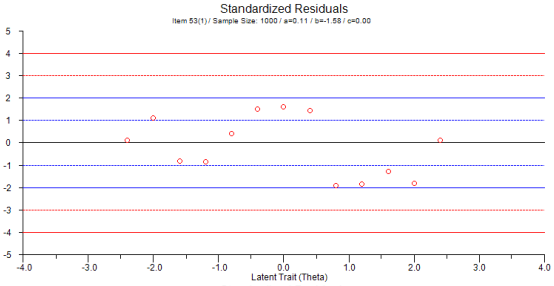

Standardized Residuals
Item 39(1) / Sample Size: 1000 / a=0.63 / b=0.19 / c=0.00
Latent Trait (Theta)

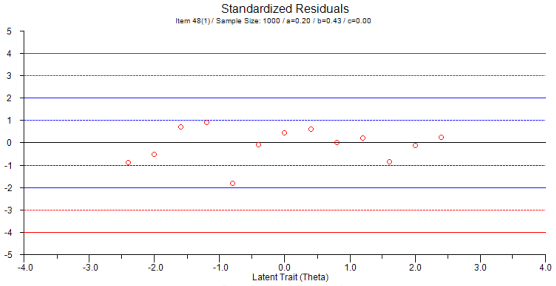

Standardized Residuals
Item 45(1) / Sample Size: 1000 / a=0.09 / b=-1.30 / c=0.00
Latent Trait (Theta)

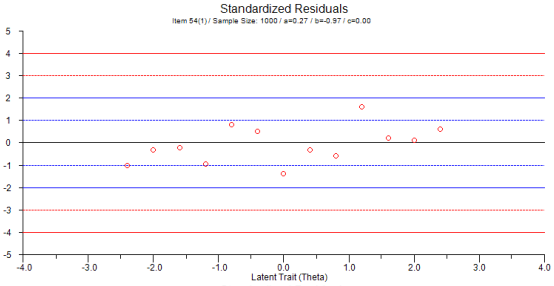

Standardized Residuals
Item 40(1) / Sample Size: 1000 / a=0.13 / b=-1.12 / c=0.00
Latent Trait (Theta)

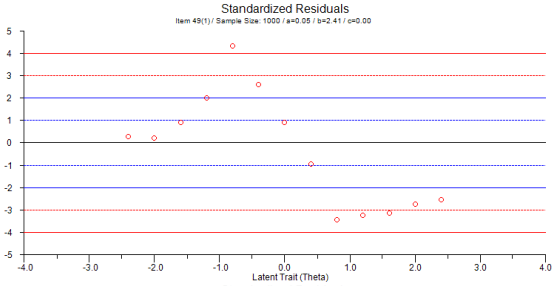

Standardized Residuals
Item 46(1) / Sample Size: 1000 / a=0.23 / b=-2.86 / c=0.00
Latent Trait (Theta)

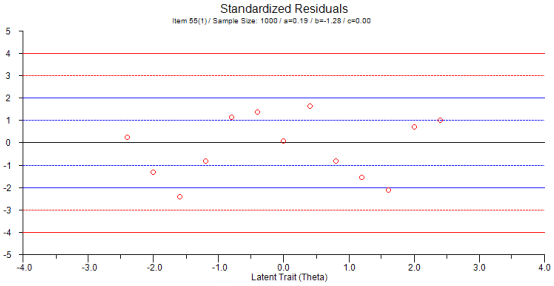

Standardized Residuals
Item 47(1) / Sample Size: 1000 / a=0.14 / b=-3.48 / c=0.00
Latent Trait (Theta)

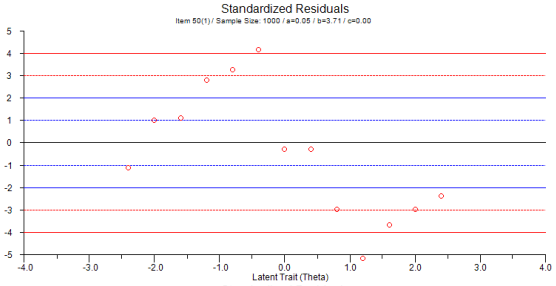

Standardized Residuals
Item 53(1) / Sample Size: 1000 / a=0.11 / b=-1.58 / c=0.00
Latent Trait (Theta)

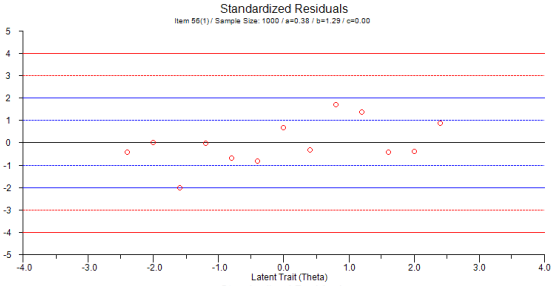

Standardized Residuals
Item 48(1) / Sample Size: 1000 / a=0.20 / b=0.43 / c=0.00
Latent Trait (Theta)

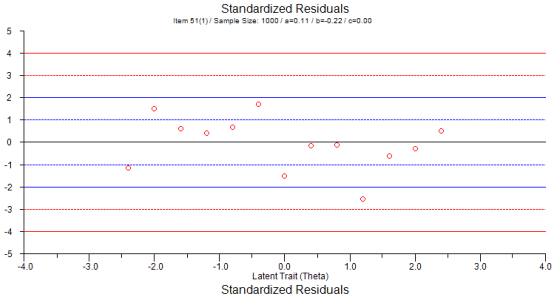

Standardized Residuals
Item 54(1) / Sample Size: 1000 / a=0.27 / b=-0.97 / c=0.00
Latent Trait (Theta)

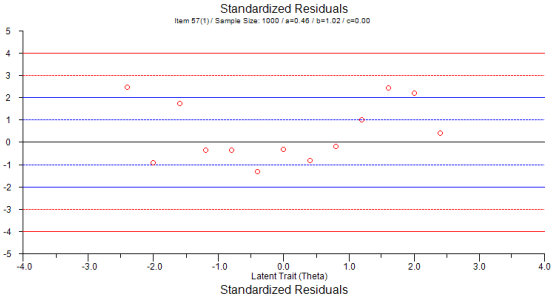

Standardized Residuals
Item 49(1) / Sample Size: 1000 / a=0.05 / b=2.41 / c=0.00
Latent Trait (Theta)

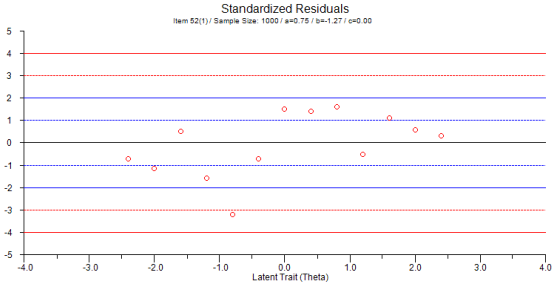

Standardized Residuals
Item 55(1) / Sample Size: 1000 / a=0.19 / b=-1.28 / c=0.00
Latent Trait (Theta)

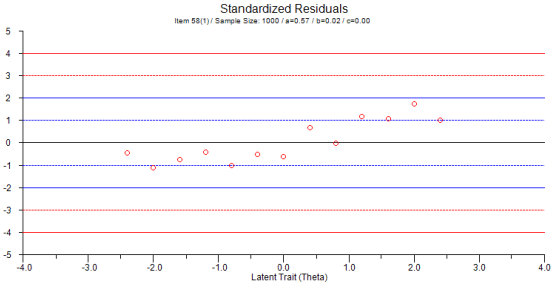

Standardized Residuals
Item 50(1) / Sample Size: 1000 / a=0.05 / b=3.71 / c=0.00
Latent Trait (Theta)

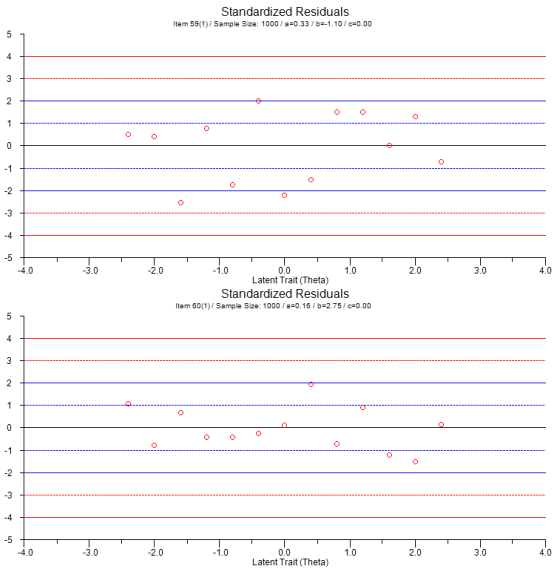

Standardized Residuals
Item 56(1) / Sample Size: 1000 / a=0.38 / b=1.29 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 51(1) / Sample Size: 1000 / a=0.11 / b=-0.22 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 57(1) / Sample Size: 1000 / a=0.46 / b=1.02 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 52(1) / Sample Size: 1000 / a=0.75 / b=-1.27 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 58(1) / Sample Size: 1000 / a=0.57 / b=0.02 / c=0.00
Latent Trait (Theta)


Standardized Residuals
Item 59(1) / Sample Size: 1000 / a=0.33 / b=-1.10 / c=0.00
Latent Trait (Theta)

Standardized Residuals
Item 60(1) / Sample Size: 1000 / a=0.16 / b=2.75 / c=0.00
Latent Trait (Theta)

Appendix F. Letter to Executive Director (MANEB)

Tamandani A Chikoko
8th July, 2013
tamandanichikoko156@gmail.com
Cell: 0993414750

The Executive Director
Malawi Nation Examination Board (MANEB)
P.O.BOX191, Zomba

Dear Sir

REQUESTING FOR PERMISSION TO USE MANEB 2009 MSCE ENGLISH PAPER1 FOR ACADEMIC RESEARCH

I write to request your permission to use the testing instrument mention above for the purpose of research study that will be conducted in schools within Zomba City.

I am a student at Chancellor College undergoing a Master of Education in Testing, Measurement and Evaluation program. This study is a partial fulfillment towards the award of my degree.

The study would be about examining the possibilities of estimating item parameters with small sample sizes in item response theory. The results of the study are expected to: reduce pretesting costs, because smaller samples would be sufficient and improve test security by reducing item exposure (fewer examinees need to see each item to estimate the item parameters accurately)

Looking forward to your favorable consideration.

Yours Faithfully

Tamandani A Chikoko (MR.)

Appendix E. Letter to the South East Education Division Manager

Tamandani A Chikoko
8th July, 2013
tamandanichikoko156@gmail.com
Cell: 0993414750

The Executive Director
Education Division Manager (SEED)
P.O.BOX148, Zomba

Dear Sir

REQUESTING FOR PERMISSION TO ADMINISTER MANEB 2009 MSCE ENGLISH PAPER1 FOR ACADEMIC RESEARCH

I write to request your permission to administer the testing instrument mentioned above to some of your schools in Zomba for purpose of academic research.

I am a student at Chancellor College undergoing a Master of Education in Testing, Measurement and Evaluation program. This study is a partial fulfillment towards the award of my degree.

The study would be about examining the possibilities of estimating item parameters with small sample sizes in item response theory. The results of the study are expected to: reduce pretesting costs, because smaller samples would be sufficient and improve test security by reducing item exposure (fewer examinees need to see each item to estimate the item parameters accurately)

Looking forward to your favorable response

Yours Faithfully

Tamandani A Chikoko (MR.)

THE MALAWI NATIONAL EXAMINATIONS BOARD

2009 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

ENGLISH LANGUAGE

Subject Number: M053/I

Friday, 16 October

Time Allowed: 1 h 10 mins
1:30 - 2:40 pm

PAPER I
(60 marks)

Instructions

1. **This paper contains 8 pages. Please check.**
2. **Before beginning** the examination, **fill in** your **Examination Number** on the answer sheets.
3. This paper has **60** Multiple Choice questions. You should answer **all** questions on the answer sheet provided.
4. **All** questions carry **one mark** each. When answering each question, show your answer on the answer sheet.

Sample Question

Choose the word **(A, B, C or D)** that gives the same meaning as the words in bold letters in the following sentence.

The actress was highly praised. Funny enough, she said that she did not like **soft soap**.
"Soft soap" means
A. acting.
B. bathing.
C. flattery.
D. comments.

The correct answer to the sample question is **C.**

The correct answer **C** to the sample question is then marked like this ■ on the answer sheet as shown below:

SAMPLE QUESTION [A][B]■[D]

1 [A][B][C][D]	26 [A][B][C][D]	51 [A][B][C][D]	76 [A][B][C][D]
2 [A][B][C][D]	27 [A][B][C][D]	52 [A][B][C][D]	77 [A][B][C][D]
3 [A][B][C][D]	28 [A][B][C][D]	53 [A][B][C][D]	78 [A][B][C][D]

PLEASE DO NOT TURN OVER UNTIL YOU ARE TOLD TO DO SO.

 Turn over

Instructions: Answer **all** questions.

Questions 1 to 25

In each of the following questions, choose the option (A, B, C or D) that best completes the sentence.

Example:

The bank is winding _____ its operations this year due to bankruptcy.

A. away
B. through
C. down
D. off

The correct answer is **C**.

1. In spite of the rains, the students insisted _____ playing the game.

A. at
B. on
C. about
D. in

2. Nurses are known to keep _____ strict ethics of their profession.

A. to
B. on
C. by
D. of

3. The sports minister congratulated the team _____ their success.

A. for
B. at
C. about
D. on

4. This nut is not screwed tight, it moves _____ when touched.

A. along
B. on
C. about
D. around

5. These days people have a liking _____ Nigerian films.

A. of
B. about
C. at
D. for

6. Mrs Banda always says that it is rude to break _____ when somebody else is talking.

A. by
B. through
C. in
D. on

7. A huge crowd turned _____ for the Cocacola Trophy final match at the Kamuzu Stadium.

A. out
B. in
C. on
D. around

8. Sungeni has signed _____ for part-time lessons.

A. off
B. in
C. on
D. up

9. She wanted to send her resignation letter by post but she decided instead to hand it _____ the following day personally.

A. in
B. to
C. on
D. over

Continued/...

10. During our last staff meeting, Mr Maluwa put _____ his ideas very clearly.

A. aside
B. across
C. through
D. at

11. The conference organizer did not permit any late comers _____ .

A. entering
B. enter
C. to enter
D. to entering

12. It is too hot in here, would you mind _____ the windows?

A. open
B. opening
C. to open
D. to opening

13. If I passed examinations, I _____ be very happy.

A. will
B. may
C. should
D. would

14. The teacher said he would teach the students provided that they _____ to redo the exercise.

A. agreed
B. agree
C. had agreed
D. would agree

15. I would never have accepted his apology if he _____ humble.

A. wouldn't be
B. wasn't
C. hadn't been
D. couldn't be

16. The problem could hardly be solved, _____?

A. could it
B. couldn't it
C. can it
D. can't it

17. The girl is too young to carry that table alone, _____?

A. is she
B. isn't she
C. is it
D. isn't it

18. Mataya didn't write to the girl, _____?

A. would he
B. wouldn't he
C. didn't he
D. did he

19. I must prepare for the geography test, _____?

A. must I
B. should I
C. mustn't I
D. shouldn't I

20. I am early today, _____?

A. aren't I
B. am I
C. isn't it
D. is it

21. Would you please _____ from smoking while the meeting is in progress?

A. avoid
B. refrain
C. stop
D. keep

Continued/...

22. We have to _____ that the message arrives in time.

A. certify
B. assure
C. ensure
D. secure

23. The watchman is always _____ faults with other people though he does not do his own work properly.

A. seeking
B. taking
C. putting
D. finding

24. The players were so far away that I couldn't _____ their faces.

A. see through
B. make out
C. see over
D. make up

25. Our mathematics teacher asked us to hang _____ while she was marking our work.

A. off
B. over
C. around
D. on

Questions 26 to 32

In each of the following questions, choose the answer (A, B, C or D) which has the same meaning as the underlined part of the given sentence.

Example:

Swimming can at times be <u>risky</u>.

A. difficult
B. exciting
C. dangerous
D. rewarding

The correct answer is **C**.

26. The manager welcomed <u>constructive</u> criticism from his audience.

A. helpful
B. serious
C. informative
D. favourable

27. Three passengers were <u>gravely</u> wounded in the bus accident.

A. fatally
B. mortally
C. superficially
D. seriously

28. They are <u>utterly</u> ignorant of the facts.

A. occasionally
B. completely
C. remotely
D. partially

29. We have had <u>cases</u> of indiscipline in the school lately.

A. occasions
B. episodes
C. incidents
D. series

30. There aren't enough employment opportunities to meet the <u>aspirations</u> of the growing number of school leavers.

A. expectations
B. education
C. abilities
D. qualifications

31. Four <u>successive</u> head teachers failed to instil discipline in the pupils.

A. unsuccessful
B. succeeding
C. continuous
D. consecutive

Continued/...

32. He was <u>at pains</u> to convince them that what happened was not planned.

A. He failed to explain it.
B. He was not feeling well and so could not explain it.
C. He was very angry to talk about it.
D. He struggled to explain it.

Questions 33 to 37

In each of the following questions, choose the option (A, B, C or D) that best describes and gives the function of the underlined phrases or clauses.

Example:

I will take the direction which <u>he has taken</u>.

A. adverb phrase, modifying "will take"
B. noun phrase, object of "will take"
C. adjective clause, qualifying "direction"
D. adverb clause, modifying "will take"

The correct answer is **C**.

33. It is likely <u>that I may visit you today</u>.

A. noun clause complement of "is"
B. adverb clause modifying "likely"
C. adjective clause qualifying "it"
D. noun clause in apposition to "it"

34. My parents moved to the village <u>where they were born</u>.

A. adverb clause modifying "moved"
B. noun clause in apposition to "village"
C. adjective clause qualifying "village"
D. noun clause object of "moved"

35. We have been standing here <u>for too long</u>.

A. adverb clause modifying "standing"
B. adjective phrase qualifying "here"
C. noun clause object of "here"
D. prepositional phrase modifying "have been standing"

36. The reporter wanted to discover <u>what the facts were</u>.

A. adjective phrase qualifying "to discover"
B. adverb clause modifying "to discover"
C. noun clause object of "discover"
D. noun clause in apposition to "to discover"

37. <u>Amused by the topic</u>, the students asked many questions.

A. noun clause in apposition to "students"
B. adjective phrase qualifying "students"
C. noun clause subject of "asked"
D. adverb clause modifying "asked"

Questions 38 to 44

In each of the following questions, choose the part of speech (A, B, C or D) that best describes each of the underlined words in the sentences.

Example:

He <u>is</u> clever.

A. adverb
B. noun
C. verb
D. preposition

The correct answer is **C**.

Continued/...

38. Japan is one of the <u>developed</u> countries.

A. verb
B. adverb
C. adjective
D. preposition

39. Don't ring me when I <u>am</u> in the hospital.

A. verb
B. conjunction
C. adverb
D. pronoun

40. Remember to sweep the room <u>while</u> I am away.

A. preposition
B. conjunction
C. adverb
D. adjective

41. Our English teacher speaks <u>too</u> fast.

A. preposition
B. adjective
C. conjunction
D. adverb

42. The minister <u>himself</u> is coming tomorrow to give you the answer.

A. noun
B. adverb
C. pronoun
D. adjective

43. The girl who is sitting <u>by</u> the window is our new head girl.

A. preposition
B. conjunction
C. adverb
D. adjective

44. When I asked for sugar for my tea, I was given a <u>little</u>.

A. noun
B. adjective
C. pronoun
D. adverb

Questions 45 – 50

In the following questions, choose the sentence (A, B, C or D) that has been correctly changed from direct to indirect speech for each of the sentences.

Example:

"I will go there," he said.

A. He says he will go there.
B. He said he goes there.
C. He said he would go there.
D. He said I will go there.

The correct answer is **C**.

45. The students said, "We are not going to watch the video this afternoon."

A. The students declared that they were not going to watch the video that afternoon.
B. The students declared that we are not going to watch the video this afternoon.
C. The students declared that they were not going to watch the video this afternoon.
D. The students declared that we are not going to watch the video that afternoon.

Continued/...

46. "I don't like spending my holidays with my aunt," she said.

A. She said that she didn't like spending her holidays with my aunt.
B. She said that she didn't like spending my holidays with her aunt.
C. She said that I don't like spending my holidays with my aunt.
D. She said that she didn't like spending her holidays with her aunt.

47. "Tawina doesn't want to leave tomorrow," he replied.

A. He replied that Tawina doesn't want to leave tomorrow.
B. He replied that Tawina didn't want to leave the following day.
C. He replied that Tawina didn't want to leave tomorrow.
D. He replied that Tawina doesn't want to leave the following day.

48. The teacher said, "Leave the room at once."

A. The teacher ordered me to leave the room at once.
B. The teacher ordered me leave the room at once.
C. The teacher ordered to leave the room at once.
D. The teacher ordered leave the room at once.

49. "I will find you at home," Maria said to Kamwana.

A. Maria told Kamwana that she would find him at home.
B. Maria told Kamwana that she will find him at home.
C. Maria told Kamwana that I will find you at home.
D. Maria told Kamwana that I would find her at home.

50. "Why didn't you go to school today?" her mother asked.

A. Her mother asked why she didn't go to school today.
B. Her mother wanted to know why she didn't go to school today.
C. Her mother wanted to know why she didn't go to school that day.
D. Her mother asked why didn't you go to school that day.

Questions 51 to 55

In each of the following questions, choose the order of adjectives (A, B, C or D) that best completes the sentences.

Example:

The school should buy a _____ machine.

A. modern, large, duplicating
B. duplicating, large, modern
C. large duplicating, modern
D. large, modern, duplicating

The correct answer is **D**.

51. He gave his daughter _____ shoes.

A. brown plastic new
B. brown new plastic
C. new plastic brown
D. new brown plastic

52. "This experiment requires _____," said the teacher.

A. large transparent water jar
B. transparent large jar water
C. water large transparent jar
D. large water transparent jar

Continued/...

53. All the poor woman would give her baby was some _____ clothes.

A. torn local cotton old
B. local torn old cotton
C. old torn cotton local
D. old torn local cotton

54. Everybody admired his _____ suit.

A. modern western beautiful
B. beautiful modern western
C. western modern beautiful
D. beautiful western modern

55. She wrote the poem on the _____ board.

A. classroom black big chalk
B. black classroom big chalk
C. big black classroom chalk
D. big classroom black chalk

Questions 56 – 60

In each of the following questions, choose the option (A, B, C or D) that best completes the sentence.

Example:

Maggie's _____ from school annoyed her parents.

A. withdraw
B. withdrawing
C. withdrawn
D. withdrawal

The correct answer is **D**.

56. _____ bathing is important, most children detest it.

A. Since
B. While
C. However
D. Nevertheless

57. His savings haven't improved, _____, they have been depleted.

A. consequently
B. on the contrary
C. still
D. however

58. He is clever, _____ he can't be trusted.

A. while
B. still
C. also
D. only

59. Not only is travelling exciting, _____ it is also educative.

A. since
B. because
C. but
D. however

60. I will let you _____ now but don't do it again.

A. down
B. out
C. up
D. off

END OF QUESTION PAPER

NB: This paper contains 8 pages.