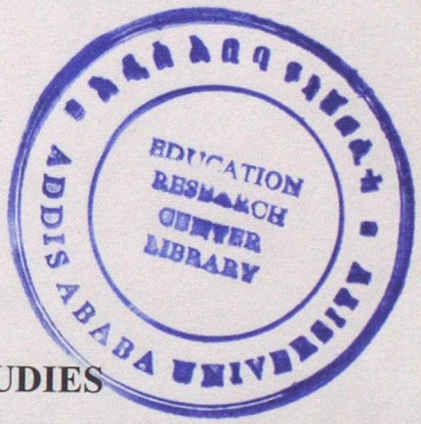
ITEM AND TEST ANALYSIS BY LANGUAGE GROUPS FOR AN EIGHTH GRADE BIOLOGY TEST IN ETHIOPIA: A COMPARISON OF IRT AND CTT MODELS

By

ZEWDU GEBREKIDAN



SCHOOL OF GRADUATE STUDIES
ADDIS ABABA UNIVERSITY

May 2011

ITEM AND TEST ANALYSIS BY LANGUAGE GROUPS FOR AN EIGHTH GRADE BIOLOGY TEST IN ETHIOPIA: A COMPARISON OF IRT AND CTT MODELS

By

ZEWDU GEBREKIDAN

SCHOOL OF GRADUATE STUDIES
ADDIS ABABA UNIVERSITY

May 2011

ITEM AND TEST ANALYSIS BY LANGUAGE GROUPS FOR AN EIGHTH GRADE BIOLOGY TEST IN ETHIOPIA: A COMPARISON OF IRT AND CTT MODELS

By

Zewdu Gebrekidan

A thesis submitted in partial fulfillment of the requirement
for the degree of
MASTER OF ARTS IN CURRICULUM AND INSTRUCTION

School of Graduate Studies
Addis Ababa University

May 2011

Acknowledgements

I would like to thank my advisor Dr Teshome Nekatibeb for inspiring me in many ways and for his continuous support throughout the project. He is my mentor and has always been very kind and supportive. My thanks also go to my organization and I am indebted to Dr Tesfaye Teshome and Ato Tamiru Zerihun for allowing me to pursue my study.

I extend my heartfelt gratitude to Ato Erkihun Desta, Ato Demoze Admasu, Ato Girma Lema, Prof. Tekeste Negash, Dr Damtew Teferra, Dr Tilahun Mengesha, Dr Marguerite Clark, Dr Benjamin Piper, Dr Gerry Shiel and Dr Gabrielle Matters. They made comments on the abstract and my choice of the topic at the very beginning, made available articles and encouraged me to work on such a challenging topic in the arcane field of psychometrics.

My wife Hemen Lema, my sons Abel and Leule Zewdu deserve many thanks for their patience and understanding not only during my study but also throughout my professional development. I thank my mother, sisters, brothers and all the family members, friends, and colleagues who were supportive in one way or the other.

Table of Contents

Acknowledgements ..... **i**

Table of Contents ..... **ii**

List of Tables ..... **v**

List of Figures ..... **vi**

List of Equations ..... **vii**

Abstract ..... **viii**

Acronyms and Abbreviations ..... **ix**

Chapter. 1 Introduction **1**

1.1 Background 3

1.2 Statement of the problem 5

1.3 Objectives 6

1.4 Research questions 7

1.5 Significance of the study 7

1.6 Delimitation and limitation of the study 8

1.7 Operational Definitions 9

Chapter. 2 Literature Review **10**

2.1 Classical Test Theory (CTT) 12

2.1.1 Assumptions of CTT 13

2.1.2 Item Parameters of CTT 14

2.2 Item Response Theory (IRT) 16

2.2.1 Assumptions of IRT 19

2.2.2 Item Response Theory Models 22

2.3 Test Translation and Test Fairness for All Groups 24

Chapter. 3 Design and Methodology **28**

3.1 Data Source and Sample Size 28

3.2 Data Analysis 29

3.2.1 Item Response Theory Methods 29

3.2.2 Classical Test Theory Methods 30

3.2.3 Statistical Packages 30

Chapter. 4 Results and Discussion **31**

4.1 IRT results for the whole group 31

4.1.1 Item Characteristic Curves for the whole group 31

4.1.2 IRT parameter estimates for the whole group 33

4.1.3 Test information curve for the whole group 34

4.1.4 Item Person Dual Plots for the whole group 34

4.1.5 One way analysis of ability score by language 36

4.1.6 Box and whisker plots 36

4.1.7 Quintals of ability score by language 37

4.1.8 Recursive partitioning analysis of IRT ability score by language 37

4.2 IRT results for Afan Oromo language group 38

4.2.1 Item Characteristic Curves for the Afan Oromo language group 39

4.2.2 IRT parameter estimates for Afan Oromo language group 40

4.2.3 Test information curve for Afan Oromo language group 41

4.2.4 Item Person Dual Plots for Afan Oromo language group 41

4.3 IRT results for English language group 42

4.3.1 Item Characteristic Curves for the English language group 43

4.3.2 Test information curve for the English language group 44

4.3.3 Item Person Dual Plots for the English language group 44

4.3.4 IRT parameter estimates for English language group 46

4.4 IRT results for Somali language group 46

4.4.1 Item Characteristic Curves for the Somali language group 47

4.4.2 Test information curve for the Somali language group 48

4.4.3 IRT parameter estimates for Somali language group 48

4.4.4 IRT parameter estimates for English language group 50

4.5 IRT results for Tigrigna language group 50

4.5.1 Item Characteristic Curves for the Tigrigna language group 51

4.5.2 Test information curve for the Tigrigna language group 52

4.5.3 Item Person Dual Plots for the Tigrigna language group 52

4.5.4 IRT parameter estimates for Tigrigna language group 54

4.6 Findings based on Classical Test Theory 55

4.7 CTT results for the whole group 55

4.8 CTT results for Afan Oromo language group 56

4.9 CTT results for English language group 57

4.10 CTT results for Somali language group 58

4.11 CTT results for Tigrigna language group 60

4.12 Summary test statistics for the whole data and by language groups 62

Chapter. 5 Summary, conclusions and recommendations 63

5.1 Summary 63

5.2 Conclusions 65

5.3 Recommendations 66

References 67

Appendix Test blue print 73

List of Tables

Table 1. Sample size by language group 28

Table 2. Sample size by region 29

Table 3. IRT parameter estimates based on the whole group 33

Table 4. Quintiles of ability score by language 37

Table 5. Analysis of variance by language 37

Table 6. IRT parameter estimates for Afan Oromo language group 40

Table 7. IRT parameter estimates for English language group 46

Table 8. IRT parameter estimates for Somali language group 50

Table 9. IRT parameter estimates for Tigrigna language group 54

Table 10.Item Analysis Results for the Whole Group 56

Table 11.Item Analysis Results for Afan Oromo Language Group 57

Table 12.Item analysis result for English language group 58

Table 13.Item analysis results for Somali language group 59

Table 14.Item Analysis Results for Tigrigna Language Group 61

Table 15. Summary test statistics based on the whole data and by language group 62

List of Figures

Figure 1. Item characteristics curve showing the relationship between the location on the latent trait and the probability of answering the item correctly 21

Figure 2. Latent trait model item plots for the whole Group 32

Figure 3. Test Information Function of the Whole Group 34

Figure 4. Item person dual plot for the whole group 35

Figure 5. Box plots for ability score by language 36

Figure 6. Recursive partitioning analysis based on IRT ability score by language 38

Figure 7. Latent trait model item plots for Afan Oromo language group 39

Figure 8. Test Information Function of Afan Oromo Language Group 41

Figure 9. Item person dual plots for Afan Oromo language group 42

Figure 10. Latent trait model item plots for English language group 43

Figure 11. Test Information Function of English Language Group 44

Figure 12. Item person dual plot for English language group 45

Figure 13. Latent trait model item plots for Somali language group 47

Figure 14. Test Information Function of Somali Language Group 48

Figure 15. Item person dual plot for Somali language group 49

Figure 16. Latent trait model item plots for Tigrigna 51

Figure 17. Test Information Function of Tigrigna Language Group 52

Figure 18. Item person dual plot for Tigrigna language group 53

List of Equations

Equation 1. Observed score based on CTT 13

Equation 2. Population mean of observed score 13

Equation 3. One parameter logistic model 22

Equation 4. Two parameters logistic model 23

Equation 5. Three parameters logistic model 23

Item and Test Analyses by Language Groups for an Eighth Grade Biology Test in Ethiopia: *A Comparison of IRT and CTT Models*

Zewdu Gebrekidan[1]

Abstract

This study carried out an analysis of item- and test-level data from the Grade 8 Biology Test of the Ethiopian Third National Learning Assessment (ETNLA). A total of 10,795 students sat for the biology test in 2007, of these 9,552 were used for the study. The test was originally prepared in English and was then translated into three language versions (Afan Oromo, Somali and Tigrigna). The main purpose was to see how the items worked across language groups. A two Parameter Logistic Model (2PLM) based on Item Response Theory was used to investigate latent traits and the main statistics generated were IRT ability scores and IRT parameter estimates (difficulty level and discrimination index). Item Characteristic Curves (ICC) and Item Person Dual Plots were generated for all 40 items by language groups. Based on the IRT ability scores, language groups were compared using one-way anova and recursive partitioning analysis. Item and test statistics were also computed following Classical Test Theory (CTT) model and results were compared with that of IRT. The Item Characteristic Curves (ICC) differed from the expected ogive shape and varied across language groups. The Test Information Function (TIF) also varied across language groups indicating the test as a whole and items in particular did not work the same way for the subgroups. A recursive partitioning analysis result based on IRT ability scores showed 20% (R^2=0.20, $F_{(3,\ 9518)}$, $p < .001$) *of the variations in achievement score was accounted by differences in language of instruction. The variance explained using CTT procedure was 13.4%* (R^2=0.134, $F_{(3,\ 9548)}$, $p < .001$). *The number of problem items (items which were too difficult and or with very low discrimination power) by language group based on CTT were: Somali (19), Afan Oromo, (12), English (10) and Tigrigna (8).The highest test score (20) was for Tigrigna, followed by Afan Oromo (18). The English language group students scored the least (15). The performance of Somali language group students were about equal to that of English group ones. The finding show that there were a number of items which did not work the same way across the four language groups which make them as language Differential Item Functioning (DIF) suspects. Based on the findings it is recommended that in the future detailed item and test analysis following the IRT model should be employed across subgroups on the pilot as well as on the operational tests. This will help to further explore DIF in future administrations of the test in order to determine whether these patterns represent real differences in achievement levels or a systematic bias that is inappropriately impacting on the scores of particular student groups.*

Key words: Biology, Item Analysis, IRT, CTT, Language DIF

---

[1] Addis Ababa University, School of Graduate Studies

Acronyms and Abbreviations

1PLM	One Parameter Logistic Model
2PLM	Two Parameters Logistic Model
3PLM	Three Parameters Logistic Model
CTT	Classical Test Theory
D	Discrimination Index
DIF	Differential Item Functioning
ETNLA	Ethiopian Third National Learning Assessment
GEQAEA	General Education Quality Assurance and Examinations Agency
ICC	Item Characteristic Curve
IRT	Item Response Theory
NLA	National Learning Assessment
p	Difficulty Level
PISA	Program for International Students Assessment
TIF	Test Information Function
TIMSS	Trends in International Mathematics and Science Studies

Item and Test Analysis by Language Groups for an Eighth Grade Biology Test in Ethiopia: *A Comparison of IRT and CTT Models*

Chapter. 1 Introduction

In Ethiopia, National Learning Assessments have been carried out at different grade levels and biology was one of the subjects tested at Grade 8. The test was originally prepared in English and translated into three languages of instruction. All the necessary measures were taken to make the test valid and reliable during its development and translation. During the last decades, test adaptations and translations have become prevalent because of an increase in national assessment testing in multiple languages, and a growing concern to test students in their first language. The comparability of test results across different language versions of these tests is at the core of the validity of interpretations in these assessments.

Test developers of large scale assessments have always tried to construct a set of items which provides an estimate of a test-takers ability and is as fair and accurate as possible to all groups of the population. As such, the test development should embrace a systematic item analysis to make sure that all the examinees with the same underlying level of knowledge have the same probability of getting an item correct (Camilli and Shephard, 1994). In other words, attempts should be made to make sure that the test and its items are not biased toward a particular group, and therefore any differences in test results are only due to the differences in the ability under measurement and not the artifacts of some other factors.

As part of the test development process, analysis of the items is a crucial part. Two prevailing methods, both with strengths and weaknesses, are predominantly used. In the Classical Test Theory (CTT), its ease of use and adaptability in analyzing practically all kinds of tests renders it a popular choice. However, its strong dependence on the kind of sampling required often limits its applicability. Hence, CTT developed tests would see the need for bigger sampling every now and then which in the long run renders it expensive. On the other hand, the emerging Item Response Theory (IRT) seems to have found a way to avoid the pitfalls of CTT. It is said to be sample free or sample independent. The only drawback is the cumbersome statistical analysis required which many test developers would shy away from. Nevertheless, IRT is slowly gaining momentum in the field of psychology (Andrade, Tavares and Valle, 2000).

When translating items both judgmental and statistical techniques should be used to ensure item comparability across languages, and rigorous quality-control steps should be included in the translation process. In this study, IRT and CTT models are used to evaluate the comparability of translated items. Language group comparison analysis is an important part of test development as it helps to examine and eliminate the items which may be potentially unfair to some groups of test takers because of language group membership. To the best of my knowledge no such study has ever been done in Ethiopia. This study may initiate others to do the same in the other subjects.

This paper in Chapter 1 started with introduction and provides the background of the study. The statements of the problem, objectives, research questions, significance of the study and operational definition are also presented under the same chapter. Chapter 2 deals with review of related literatures focusing on the two Test Theories

and Chapter 3 deals with methodological issues. The findings of the study are presented under Chapter 4. Chapter 5 presents summary conclusions, and recommendations.

1.1 Background

A test can be studied from different angles and the items in the test can be evaluated according to different theories. Two such theories are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970's IRT has more or less replaced the role CTT had and is now the major theoretical framework used in this scientific field (Crocker and Algina, 1986; Hambleton and Rogers, 1990; Hambleton, Swaminathan, and Rogers, 1991).

With the emerging trend of developing local instruments more and more research has been developed in producing achievement and psychological tests. Most often, these researchers rely on Classical Test Theory (CTT) to develop these instruments in spite of the strong presence of Item Response Theory in the recent decades.

In CTT, test scores are said to be composed of three components: test score, true score, and error score. The invariance is brought about by differences contributed by the sample from which the scores were derived. Again, here lies the dependence of CTT on the sample the scores were taken from. However, IRT addresses this by disregarding the sample and instead looking at the characteristics of the item or item parameters. By focusing on the items, the issue of sampling becomes negligible. One can now generalize better item-generated scores across samples and person abilities (Hambleton, Swaminathan, and Rogers, 1991).

Item response theory came about to rectify some of the short comings of classical test theory. In classical test theory, a student's ability is determined by the score on a particular test. The difficulty and discrimination of a particular item as well as the reliability and validity of the test are determined by the ability of a group of students; if the characteristics of the group changes so do these factors. This makes it difficult then to compare students across different tests and to compare items across different student groups. Additionally if the students who take a particular test are of different ability, then their scores will have different amounts of error, which is contrary to the assumption behind the standard error of measurement that it is the same for all individuals. In developing a test, it would be preferable to be able to predict how different groups or individuals will perform on a given item. In classical test theory the emphasis is on the test, while in IRT the emphasis is on the items.

The underlying assumption of IRT is that the responses on a particular test are accounted for by a small number of latent traits. As Crocker and Algina write

> At the heart of the theory is a mathematical model of how examinees at different ability levels for the traits would respond to an item. This knowledge allows one to compare the performance of examinees who have taken different tests. It also permits one to apply the results of an item analysis to groups with different ability levels than the group used for the item analysis (Algina and Crocker, 1986, p. 109).

IRT is useful in building tests, identifying potentially biased test items, equating scores from different tests or forms of the same test, developing tests which can discriminate at a particular level of ability, and in the development of tailored testing systems.

Studies linking CTT and IRT item characteristics have been done and have shown signs of positive indications of a relationship that exists (Adedoyin, Nenty, and Chilisa, 2008; Nukhet, 2002; Fan, 1998). It is then, the goal of this paper to analyze the item characteristics of a Grade 8 biology test using both CTT and IRT methods and to check if the methods are comparable and can be used independently or interchangeably.

National Learning Assessment (NLA) in Ethiopia is a large scale national survey conducted by an agency under the Ministry of Education. It was first conducted in 2000 and has been repeated every three or four years since then. The present study aimed to investigate item and test analysis of the biology test of the Ethiopian Third National Learning Assessment (ETNLA). A total of 10,795 students sat for the biology test in 2007. The test was originally prepared in English and was then translated into three language versions namely Afan Oromo, Somali and Tigrigna (GEQAEA, 2007). The effect of these test languages will be investigated.

1.2 Statement of the problem

Research studies about the psychometric characteristics of translated test items indicate more than the expected amount of differential item functioning (DIF) on different language versions of the same test, even though some strict translation processes are followed during test adaptation. For example, Gierl, Rogers, and Klinger (1999) have reported that 52% of items on a Canadian achievement test displayed DIF across English and French speaking examinees. Allalouf, Hambleton, and Sireci (1999) have reported that 34% of verbal items on the Israeli Psychometric Entrance Test displayed DIF across Hebrew and Russian examinees. Arim and Ercikan (2005) have reported that 23% of items on the Third International

Mathematics and Science Study–Repeat (TIMSS-R) displayed DIF when English and Turkish speaking examinees were compared. In Ethiopia the variance explained in biology scores due to test language difference in the ETNLA was the highest (13.8%) when compared with the other subjects: mathematics (2.3%), chemistry (5.7%) and physics (5.9%) (GEQAEA, 2007).

In this situation, it is crucial to make sure that test items have been scrutinized for possible bias because test bias is a threat to valid interpretation of test scores (Langenfeld, 1997). Therefore, screening items for differential item functioning should become part of test scrutiny prior to administration of the tests since differential item functioning is a necessary condition for item bias (Clauser and Mazor, 1998). The study focuses on item and test analysis following IRT and CTT models to identify problem items across test languages.

1.3 Objectives

The main objective of the study is to carry out item and test analysis on Grade 8 Biology test of the Ethiopian Third National Learning Assessment (ETNLA) using IRT and CTT procedures. The specific objectives are to:

- determine the difficulty level and discrimination index of each item based on item analysis following IRT and CTT models,
- compare the item and test statistics obtained using the two models,
- make comparisons across language groups using IRT ability scores and
- identify items which did not work as expected across the four language groups.

1.4 Research questions

The study will be guided by the following research questions.

1. What are the difficulty level and the discrimination index of each item for the whole group?

2. What are the difficulty level and the discrimination index of each item for each language group?

3. Are there any biology items that function differentially between the four language groups?

1.5 Significance of the study

National Learning Assessment in Ethiopia is the only large scale study which provides information on the quality of education and the factors associated. The achievement scores are taken as major indicator of quality. Hence further analysis will provide additional information and will help future improvements in the test development process. To ensure that translated items are equivalent to their original versions, both statistical and qualitative analysis are necessary.

Item Analysis using Classical Test Theory (CTT) was carried out during pilot testing for the purpose of improving the items. The English version test was translated into the respective test languages and equivalencies were checked through independent back translation and experts' judgment. Nevertheless, detailed item level analysis in general and IRT model in particular had never been conducted on the pilot and final tests. This study is intended to fill this gap and may create an opportunity to use the wealth of data available. A two-parameter IRT model will be used to generate ability

score, and estimate parameters. By evaluating translated items statistically, test developers can ensure the comparability of tests across languages and they can identify the types of problems that should be avoided in future translation efforts.

1.6 Delimitation and limitation of the study

The scope of this study is the response of the sample students who sat for the biology tests in the Ethiopian Third National Learning Assessment disaggregated by the four test languages.

There are varieties of IRT methods; however, this study is limited to the use of a two parameter logistic IRT model. Using multiple models and making comparisons of the findings would have been the better option but the scope of the study, time and budget constraints did not allowed.

1.7 Operational Definitions

The definitions presented here are based on Camilli and Shepard (1994) and Clauser and Mazor (1998).

Item analysis: A set of statistical techniques to examine the performance of individual items. This is important when developing a test or when adopting a known measure.

DIF: DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure.

Item bias: Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias.

Chapter. 2 Literature Review

Test developers are basically concerned about the quality of test items and how examinees respond to it when constructing tests. A psychometrican generally uses psychometric techniques to determine the validity and reliability. Psychometric theory offers two approaches in analyzing test data namely Classical Test Theory (CTT) and Item Response Theory (IRT). Both theories enable to predict outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers. Both are concerned to improve the reliability and validity of psychological tests. Both of these approaches provide measures of validity and reliability. There are some identified issues in the classical test theory that concerns with calibration of item difficulty, sample dependence of coefficient measures, and estimates of measurement error which in turn is addressed by the item response theory.

CTT has dominated the area of standardized testing and is based on the assumption that a test-taker has an observed score and a true score. The observed score of a test-taker is usually seen as an estimate of the true scores of that test-taker plus/minus some unobservable measurement error (Crocker and Algina, 1986; Hambleton and Swaminathan, 1985). An advantage with CTT is that it relies on weak assumptions and is relatively easy to interpret. However, CTT can be criticized since the true score is not an absolute characteristic of a test-taker since it depends on the content of the test. If there are test-takers with different ability levels a simple or more difficult test would result in different scores. Another criticism is that the items' difficulty could vary depending on the sample of test-takers that take a specific test. Therefore, it is difficult to compare test-takers' results between different tests. In the end, good

techniques are needed to correct for errors of measurement (Hambleton, Robin, and Xing, 2000).

IRT was originally developed in order to overcome the problems with CTT. A major part concerning the theoretical work was produced in the 1960's (Birnbaum, 1968; Lord and Novick, 1968) but the development of IRT continues (van der Linden and Glas, 2000). One of the basic assumptions in IRT is that the latent ability of a test-taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test-taker can be modeled in different ways depending on the nature of the test (Hambleton et al., 1991). It is common to assume unidimensionality, that is, the items in a test measure one single latent ability. According to IRT, test-taker with high ability should have a high probability of answering an item correctly. Another assumption is that it does not matter which items are used in order to estimate the test-takers' ability. This assumption makes it possible to compare test-takers' result despite the fact that they have taken different versions of a test (Hambleton and Swaminathan, 1985). IRT has been the preferred method in standardized testing since the development of computer programs. The computer programs can now perform the complicated calculations that IRT requires (van der Linden and Glas, 2000). There have been studies that compare the indices of CTT and IRT (Bechger, Gunter, Huub, and Beguin, 2003). Other studies have aimed to compare the indices and the applicability of CTT and IRT, see for example how it is use in the Swedish Scholastic Aptitude Test in Stage (2003) or how it can be used in test development (Hambleton and Jones, 1993).

2.1 Classical Test Theory (CTT)

Classical test theory is regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation inability of interest. All other potential sources of variation existing in the testing materials such as external conditions or internal conditions of examinees are assumed either to be constant through rigorous standardization or to have an effect that is nonsystematic or random by nature (Van der Linden and Hambleton, 2004). The central model of the classical test theory is that observed test scores are composed of a true score and an error score where the true and the error scores are independent. The variables are established by Spearman (1904) and Novick (1966).

Traditionally, methods of analysis based on classical test theory have been used to evaluate tests. The focus of the analysis is on the total test score; frequency of correct responses (to indicate question of difficulty); frequency of responses (to examine distracters); reliability of the test and item-total correlation (to evaluate discrimination at the item level) (Impara and Plake, 1997). Although these statistics have been widely used, one limitation is that they relate to the sample under scrutiny and thus all the statistics that describe items and questions are sample dependent (Hambelton, 2000).

A test theory and test model is a symbolic representation of the factors influencing the observed test scores and is described by its assumptions. Classical test theory describes how errors of measurement can influence the observed scores of a test. An observed score is expressed as the sum of the true score and the error of measurement. It is this central idea of the relationship among true score, observed score and error of

measurement that enables the classical test theory to describe the factors which influence the test scores.

2.1.1 Assumptions of CTT

The classical true score theory is underpinned by seven assumptions (Yen and Allen., 1979, pp.57-60). These seven assumptions are stated below.

Assumption one states that an observed score *(X)* in a test is the sum of two parts known as (1) the true score (*T*) and (2) the error score (*E*) or error of measurement. Mathematically, this assumption is expressed as (Equation 1):

$$X=T+E$$

Equation 1. Observed score based on CTT

The additive nature of the true score and the error score is commonly made in statistical work, because it is mathematically simple and appears reasonable.

Assumption two states that the expected value (ξ) or population mean of an observed score is the true score. Mathematically, this assumption is expressed as:

$$X(\xi) = T$$

Equation 2. Population mean of observed score

Equation 2 defines the true score as the mean of the theoretical distribution of the observed scores that would be found in repeated independent testing of the same person with the same test. The true score is viewed as remaining constant over all administrations, and over all parallel forms of a test. Algina and Crocker (1986, p. 109) define the true score as the mean or expected value of a random variable.

Assumption three states that the error scores and the true scores obtained by a population of examinees on one test are uncorrelated.

Assumption four states that the error scores on two different tests are uncorrelated.

Assumption five states that the error scores on one test are uncorrelated with the true scores on another test.

Assumption six states the definition of parallel tests. If observed score, true score and error variance of test *A* and observed score, true score and error variance of test *B* are the same then test *A* and *B* are parallel tests.

Assumption seven states that the tests that are essentially equivalent have true scores that are the same except for an additive constant.

2.1.2 Item Parameters of CTT

The quality of items is judged based on their parameters like item difficulty, item discrimination, item reliability and item validity statistics. These item parameters help test makers to choose the right items in accordance with the construct of interest when making a test.

2.1.2.1 Item Difficulty

Item difficulty (sometime known as item facility) is the proportion of examinees who answer an item correctly (Algina and Crocker 1986, p.90). Item difficulty in the context of CTT is sample dependent. Its values will remain invariant only for groups of examinees with similar levels. Item difficulty is often referred to as p-value in CTT. This value represents the percentage of a certain group of examinees who selected a particular response. A p-value can be calculated for each response, the

correct answer and each of the distractors, by dividing the number of individuals that selected a particular response by the total number of individuals in the group of interest.

The p-values of items will be different for different items of a test depending on the types of examinees. If the items are difficult, then their p-values will be low. If the items are easy, then p-values will be high.

The p-value of an item can provide general guidance when analyzing an item. If the p-value is very low (in the range of 0.00 to 0.20), then the item is very hard and the possibility that the item has been wrongly keyed or that there is more than one correct answer to the question should be examined. Very low p-value is also indicative of floor effect.

If the p-value is greater than 0.95, then the correct answer is probably too obvious for the test population. The very high p-value is also indicative of ceiling effect. The items with p-values less than or equal to 0.20 and greater than or equal to 0.95 should be deleted or revised to present a greater challenge to the test candidates. If the p-value is zero for any response, this is called a "Null distracter." Null distracters are indicative of obvious answers, nonparallel distracters, or nonsensical distracters.

2.1.2.2 Item Discrimination

Examinees differ in their abilities. It is conventional to expect high scores, average scores and low scores and other scores which incline to fall in any of these groups. Therefore, while analyzing test items, one of the objects is to select items which have potential to separate examinees into different categories of performance based on their abilities. This means that a test item should have characteristics capable of being

scored correctly by high ability examinees and incorrectly by low ability examinees. The items which have such properties are discriminative. These items discriminate examinees who know answers from examinees who do not know answers.

Index of item discrimination (D) is applicable only to dichotomously scored items. To calculate item discrimination index, examinees are separated into two groups based on their total test scores with respect to the cut scores. The two groups are categorized as upper group and lower group.

Algina and Crocker (1986, p.315) and (Ebel, 1965) provide the following guidelines for interpretation of D-values when the groups are established with total test score as the criterion: If $D \geq 0.40$, the item is functioning quite satisfactorily. If, $0.30 \leq D \leq 0.39$ little or no revision is required. If, $0.20 \leq D \leq 0.29$ the item is marginal and needs revision. If, $D \leq 0.19$ the item should be eliminated or completely revised.

2.2 Item Response Theory (IRT)

Another branch of psychometric theory is the item response theory (IRT). IRT may be regarded as roughly synonymous with latent trait theory. It is sometimes referred to as the strong true score theory or modern mental test theory because IRT is a more recent body of theory and makes stronger assumptions as compared to classical test theory. This approach to testing based on item analysis considers the chance of getting particular items right or wrong. In this approach, each item on a test has its own item characteristic curve that describes the probability of getting each particular item right or wrong given the ability of the test takers (Kaplan and Saccuzzo, 1997).

The Rasch model as an example of IRT is appropriate for modeling dichotomous responses and models the probability of an individual's correct response on a

dichotomous item. The logistic item characteristic curve, a function of ability, forms the boundary between the probability areas of answering an item incorrectly and answering the item correctly. This one-parameter logistic model assumes that the discriminations of all items are assumed to be equal to one (Maier, 2001). Another fundamental feature of this theory is that item performance is related to the estimated amount of respondent's latent trait (Anastasi and Urbina, 2002). In cognitive tests, latent traits are called the ability measured by the test. The total score on a test is taken as an estimate of that ability. A person's specified ability succeeds on an item of specified difficulty.

The benefit of the item response theory is that its treatment of reliability and error of measurement through item information function are computed for each item (Lord, 1980). These functions provide a sound basis for choosing items in test construction. The item information function takes all items parameters into account and shows the measurement efficiency of the item at different ability levels. Another advantage of the item response theory is the invariance of item parameters which pertains to the sample-free nature of its results. In the theory the item parameters are invariant when computed in groups of different abilities. This means that a uniform scale of measurement can be provided for use in different groups. It also means that groups as well as individuals can be tested with a different set of items, appropriate to their ability levels and their scores will be directly comparable (Anastasi and Urbina, 2002).

Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of a positive response. A variety of IRT models have been developed for dichotomous and polytomous data. In

each case, the probability of answering correctly or endorsing a particular response category can be represented graphically by an item (option) response function (IRF/ORF). These functions represent the nonlinear regression of a response probability on a latent trait, such as conscientiousness or verbal ability (Hulin, Drasgow, and Parsons, 1983).

IRT provides several advantages over classical test theory (CTT) methods for constructing tests and examining measurement equivalence. Unlike CTT item statistics, which depend fundamentally on the subset of items and persons examined, IRT item and person parameters are invariant. This makes it possible to examine the contribution of items individually as they are added and removed from a test. Moreover, IRT allow researchers to calculate conditional standard errors of measurement based on a test information function, rather than assuming an average standard error across all trait levels as in CTT. This allows researchers to select items that provide maximum measurement precision in a particular ability/trait range (Hulin *et al.*, 1983).

Second, IRT allows researchers to conduct rigorous tests of measurement equivalence across experimental groups. This is particularly important in cross-cultural research where groups are expected to show mean differences on the attribute being measured. IRT methods can distinguish item bias from true differences on the attribute measured, whereas CTT methods cannot (Kim, Cohen, and Park, 1995).

IRT also facilitates computer adaptive testing. Items can be selected that provide the most information for each examinee. This can dramatically reduce time and costs associated with test administration (Hulin et al., 1983).

Item response theory postulates that (a) an examinee test performance can be predicted (or explained) by a set of factors called traits, latent traits, or abilities, and (b) the relationship between an examinee item performance and the set of traits assumed to be influencing item performance can be described by a monotonically increasing function called an item characteristic function (Lord and Novick, 1968, p.359). Inherent in these theories are (a) an examinee test performance which is observable and (b) the unobservable traits or abilities assumed to underlie examinee performance on the test.

The relationship between observable test performance (responses) and unobservable traits underlying the test performance can be expressed as a mathematical function and this makes it possible to build mathematical model called item response model. Depending on the types of assumptions underlying the item response models, different types of models can be built. For instance, one-parameter logistic models, two-parameter logistic models and three-parameter logistic models have different assumptions.

2.2.1 Assumptions of IRT

Four common assumptions of Item Response Theory are (a) dimensionality of latent space, (b) local independence, (c) item characteristic curves and (d) speededness.

2.2.1.1 Dimensionality

Item response theory assumes that a set of k latent traits or abilities underlie examinee performance on a set of test items. The k latent traits define a k dimensional latent space, with each examinee's location in the latent space being determined by the examinee's position on each latent trait (Hambleton and Swaminathan, 1985, p.16).

Some item response models assume single latent trait to be sufficient to explain for examinee test performance and they are called unidimensional item response models. For unidimensionality latent space to be met adequately by a set of test data, a dominant component or factor is assumed to influence the examinee test performance (Hambleton and Swaminathan, 1985, p.17).

Some item response models assume more than a single latent trait necessary to explain for examinee test performance and they are called multidimensional item response models. The multidimensional item response models will not be described in the thesis.

2.2.1.2 Local Independence

The local independence assumption states that an examinee's responses to different items in a test are statistically independent when abilities influencing test performance are held constant (Hambleton and Swaminathan, 1985, p.23). Local independence, by definition, will hold only when the items are not related to each other (a) by content and (b) when responses to the items are not linked by clues.

2.2.1.3 Item Characteristic Curve

An item characteristic curve (ICC) is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it (Hambleton and Swaminathan, 1985, p.25).

Lord and Novick (1968, p.360) note that item characteristic function or curve remains invariant from one group of examinees to the next, resulting in the invariance of item parameters involved in generating the item characteristic curve. This is an important aspect of the item response theory which distinguishes it from the classical test theory.

Hambleton and Swaminathan, (1985, p. 18) state that the invariance of item and ability parameters mean that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items (Hambleton and Swaminathan, 1985, p.18). IRT methods model the probability of an individual's response to an item. The relationship between the probability of success to an item and the latent trait (the ability) is described by a function called item characteristic curve (ICC) that takes an S-shape (ogive) (Figure 1).

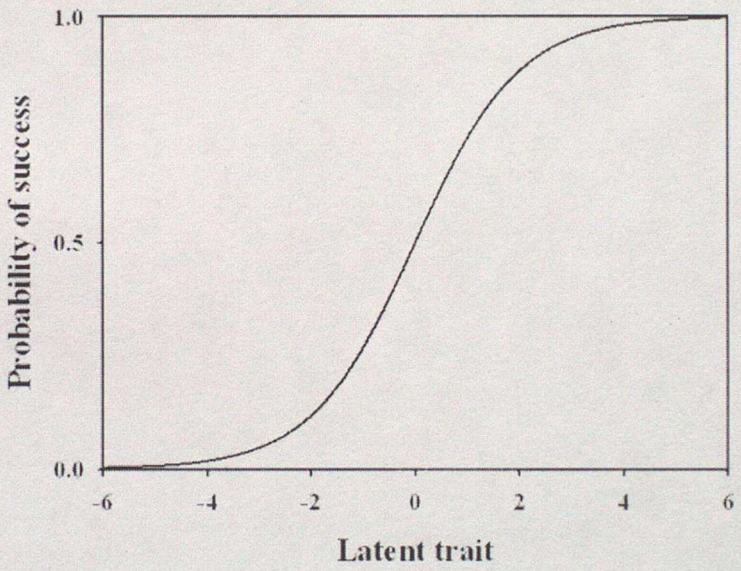


Figure 1. Item characteristics curve showing the relationship between the location on the latent trait and the probability of answering the item correctly

Source: Modified from Baker (2001)

2.2.1.4 Speededness

An implicit assumption of all commonly used IRT model is that the test to which the model fits are not administered under speeded conditions (Hambleton and Swaminathan, 1985, p.30). This assumption requires the examinees to be provided

with sufficient time to answer the items to ensure that the failure to answer test items is only because of the limited ability and not because of lack of time to answer the items.

2.2.2 Item Response Theory Models

The item difficulty statistic is an appropriate choice for achievement or aptitude tests when the items are scored dichotomously (i.e., they are either correct or incorrect). Several methods were developed by G. Rasch and A. Birnbaum to estimate test reliability.

2.2.2.1 The One Parameter Logistic Model (1PLM)

The One-Parameter Logistic Model (1PLM) is

$$P_i(B_i) = 1 / (1 + exp[-(B_i - D_i)])$$

Equation 3. One parameter logistic model

Where

B_i is the parameter describing the ability of the person being tested,

P_i is the probability of getting a correct response, and

D_i is the parameter describing the difficulty of item.

2.2.2.2 The Two Parameter Logistic Model (2PLM)

The Two-Parameter Logistic model uses an item response theory (IRT) model that specifies the probability of a correct response as a logistic distribution in which items vary in terms of their difficulty and discrimination. It is typically applied to multiple choice or short constructed-response items that are scored either correct or incorrect, and do not appear to allow for guessing.

The 2PLM (Birnbaum, 1968), generalizes the one-parameter logistic, or Rasch, model by allowing items to vary not only in terms of their difficulty (*b*) but also in terms of their ability to discriminate (*a*) among individuals of various proficiency. As with the Rasch model, the 2PL assumes that the probability of a correct guess is zero. The fundamental equation of the 2PL is the probability that a person whose proficiency on scale *k* is characterized by the unobservable variable θ_k will respond correctly to item *j*:

$$P(x_j = 1 | \theta_k, a_j, b_j) = \frac{1}{1 + e^{-Da_j(\theta_k - b_j)}}$$

Equation 4. Two parameters logistic model

where

x_j = the response to item *j*, 1 if correct and 0 if not;

a_j = the slope of parameter of item *j*, characterizing its sensitivity to proficiency, where $a_j > 0$;

b_j = the threshold parameter of item *j*, characterizing its difficulty;

D = an arbitrary scaling constant typically set to 1.7 to approximate results from the normal ogive model.

2.2.2.3 The Three Parameter Logistic Model (3PLM)

The Three-Parameter Logistic Model (3PLM, or Birnbaum's Model):

$$P_i(B_i) = C_i + (1 - C_i) / (1 + \exp[-K_a(B_i - D_i)]).$$

Equation 5. Three parameters logistic model

The parameter C_i is used when item i is constructed so that guessing the correct answer is possible.

2.3 Test Translation and Test Fairness for All Groups

One of the important factors which should be taken into account in dealing with the validity of any test is the issue of fairness (Thissen, 2001). Educational Testing Service Fairness Review Guidelines (2003) offers a simple and fairly straightforward verdict: "A test that shows valid differences is fair; a test that shows invalid differences is not fair". Therefore, when people with similar abilities in the construct being measured perform substantially differently on a test item it is necessary that the item be reviewed for fairness (Gierl, Khaliq and Boughton, 1999),and perhaps removed if the differential performance of the examinees is not balanced or cancelled over the test as a whole (Wainer, Sireci and Thissen, 1991).This issue has made a great area of investigation for researchers, and a technique which empirically measures the differential functioning of items for groups which are matched on the ability of interest has now become "the new standard in psychometric bias analysis" (Zumbo, 1999) and is now "a key component of validity studies in virtually all large-scale assessments" (Penfield and Camilli, 2007). This technique is called differential item functioning (DIF) after Holland and Thayer who used the term in a seminal chapter on test validity (1988).

DIF analysis is often used to examine group differences between specific racial or ethnic groups or between males and females. For example, Hauser and Kingsbury (2004) explored differential functioning across student groups formed based on ethnicity and based on gender on items from the Idaho Standards Achievement Test. Zenisky, Hambleton, and Robin (2004) explored gender DIF in a large-scale science assessment. Other research has also examined incidences of DIF for limited English proficient students (Snetzler and Qualls, 2000). DIF analysis have also been

conducted for students with disabilities. Specifically, DIF analysis have been used to examine effects of accommodations that are provided to students with disabilities during testing (Bolt, 2004; Cohen, Gregg, and Deng, 2005; Koretzand Hamilton, 1999).

In studies of construct validity, the assessment of DIF is an important issue in evaluating the inferences from educational and psychological testing. DIF occurs when test items display different statistical or psychometric properties in different groups of examinees after matching on the same intended-to-be-measured underlying proficiency, In other words, the absence of differential item functioning occurs when the item responses and group variables are independent after conditioning on ability (Millsap and Everson, 1993). Although the assessment of test fairness is often based on the single-item DIF analysis, DIF can occur at the subtest and test score level (Stout, 2002).

Eliminating measurement bias is important because measurement bias is a threat to validity of interpretation and use of educational measures. Issues of construct-related evidence of validity and issues of measurement bias are interrelated in the sense that the number of constructs being measured by the test or item (Ackerman, 1992). If a test lacks construct-related evidence of validity, it means that the test contains items that are measuring constructs other than those are intended to be measured, indicating that there is a potential for bias against or for a certain group of examinees.

Studies concerning test bias examine whether there are individual items that function differently among groups within the same level of ability being measured. Differential item functioning (DIF) studies are conducted to investigate these problems. Dorans

and Holland (1993) defined differential item functioning as "differences in item functioning after groups have been matched with respect to the ability or attribute that the item purportedly measures". Similarly, Angoff (1993) noted that DIF refers to "the simple observation that an item displays different statistical properties in different groups setting, after controlling for differences in the abilities of the groups...". Camilli and Shepard (1994) stated that "An item is said to be measuring differentially if examinees of the same ability do not have the same probability of answering the item correctly".

The theory underlying the development of traditional normative assessments (CTT) differs from the theory underlying the new kinds of tests based on Rasch and item response testing models (Clark and Watson, 1995; Croker and Algina, 1986; Embretson,1999; Embretson and Reise, 2000; Lord and Novick, 1968; MacDonald, 1999). Classical tests are based on individual differences and a normative interpretation in which a person's test scores are represented in terms of how others performed on the test. Scores might be expressed as standard scores which are based on distances from the mean of the normative group. The newer kinds of tests are based on latent trait theory (also known as item response theory). This approach represents a different perspective on scores and their interpretation than classical testing. Rather than comparing tests results to the performance of a normative group, scores are interpreted in terms of a continuum of the trait being assessed (Embretson and Reise, 2000).

When groups of people, males and females for example, respond differently to a test item, that item is said to be biased. Test bias can occur when groups within a target population systematically respond differentially to the test. In such cases, the validity

of the item can become a concern since it is assumed that the item measures the same trait or behavior in all people, and that the item and test have the same predictive value regardless of the groups (racial, ethnic, gender, etc.) to which they belong. Further, the definition of bias and identification of groups within a population for whom the test is biased can be difficult (Nunn ally and Bernstein, 1994). Classical test theory provides a number of ways of detecting and dealing with item bias among them, using linear regression of test scores on some criterion for the standard group and the group that might experience bias, and studying residuals to find extraneous variables that may lead to bias (Nunnally and Bernstein,1994).

The differential item function (DIF) is at the core of the IRT approach to test fairness. "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers, 1991, p.110). One approach to DIF would be to compare the item parameters from the ICC obtained by administering the test to the groups of interest. If the item parameters are the same, the test would be considered unbiased for these groups. This approach requires putting the results of the testing of the two groups on a common scale, and comparing the parameters. When the parameters adjusted to the common scale are not different, no DIF exists and the test can be considered fair for the subgroups of interest.

Chapter. 3 Design and Methodology

3.1 Data Source and Sample Size

The data used in the study are the responses of Grade 8 students to the Ethiopian Third National Learning Assessment biology items (Appendix). The test was composed of 40 multiple choice items. Each item is composed of four choices. The data were organized by language of instruction.

Table 1. Sample size by language group

Language	N	Percent
Afan Oromo	2118	22.2
English	5919	62.1
Somali	498	5.2
Tigirigna	998	10.5
Total	**9533**	**100.0**

A total of 10,795 students sat for the biology test in 2007 of these 9,552 were used for the study. The test was originally prepared in English and translated into three language versions (Afan Oromo, Somali and Tigrigna). Data from the Amhara region was excluded from the study since it shifted medium of instruction during the period of data collection (Table 1 and Table 2).

Table 2. Sample size by region

Region	N	Percent
Addis Ababa	998	10.5
Afar	891	9.3
Benshangul Gumuz	789	8.3
Dire Dawa	646	6.8
Gambella	903	9.5
Harari	682	7.2
Oromia	2055	21.6
SNNPR	1073	11.3
Somali	498	5.2
Tigray	998	10.5
Total	**9533**	**100.0**

3.2 Data Analysis

Tests can be studied from different angles and the items in the test can be evaluated according to different theories. Two such theories used in this study were Classical Test Theory (CTT) and Item Response Theory (IRT). CTT was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970's IRT has more or less replaced the role CTT had and is now the major theoretical framework used in this scientific field (Crocker and Algina, 1986; Hambleton and Rogers, 1990; Hambleton, Swaminathan, and Rogers, 1991).

3.2.1 Item Response Theory Methods

In carrying out item analysis there are several different IRT models to choose from (Thissen and Steinberg, 1986). The first consideration when choosing the right model involves the number of item response categories, as this obviously limits the choice of appropriate models. For dichotomous items, the 1, 2, and 3 parameter logistic models are available (1PLM, 2PLM, 3PLM). Since all the items of ETNLA biology test are

multiple choice and marked as right-wrong (dichotomously) the 2 parameter logistic model (2PLM) is chosen.

The main statistics generated were IRT ability scores, difficulty levels and discrimination indexes. A one way analysis of variance was employed to make comparisons across language groups using the IRT ability score as dependent variable. Further comparisons on the same score were made following recursive partitioning analysis.

Latent trait model item plots, test information function, parameter estimates and item person dual plots were produced based on the whole group and by test language.

3.2.2 Classical Test Theory Methods

A summary descriptive statistics which show mean, standard deviation of the test scores difficulty and discrimination indexes were computed across language groups. The reliability was computed with coefficient alpha, defined as using the proportion who answered the item correctly, *p*-values, and point biserial correlation, r_{pbis}. The point biserial correlation is the correlation between the test-takers' performance on one item compared to the test-takers' performances on the total test score (Crocker and Algina, 1986). In this study coefficient alpha, *p* values and r_{pbis} are used from the CTT.

3.2.3 Statistical Packages

Data organization was carried using SPSS v19 and MS Excel 2010 and analysis were carried out using Systat v13 and SAS v9.1.

Chapter. 4 Results and Discussion

The data analysis was carried out following IRT and CTT procedures by language group. Initially analysis was carried out using the whole data set then each language group is addressed separately. The findings are presented for each procedure.

IRT is also called latent trait theory because the theory assumes the existence of a latent trait, which is a tester characteristic that leads to a consistent performance on a test. The findings based on IRT procedure are presented for the whole group followed by each language group one by one. The latent trait model item plots are the main out puts and they are presented visually. The horizontal axis of the plots denotes the latent trait, the vertical axis is the proportion of examines responding correctly, and the curve is called the item characteristic curve (ICC). According to IRT, if the item is well-written, the ICC should be a normal ogive the shape of the curve looks like that of a half bell-curve.

4.1 IRT results for the whole group

In all the four language groups together 9,533 cases were analyzed; each containing 40 items, 11 cases were deleted by editing for missing data or for zero or perfect total scores after item editing. No item was deleted by editing for missing data or for zero or perfect total scores after item editing. Data below are based on 9,522 cases and 40 Items. The total score mean is 15.4 with a standard deviation of 5.49.

4.1.1 Item Characteristic Curves for the whole group

IRT can be used to create a unique plot for each item (the Item Characteristic Curve - ICC). The ICC is a plot of Probability that the Item will be answered correctly against Ability. The shape of the ICC reflects the influence of the two factors: Increasing the

difficulty of an item causes the curve to shift right - as candidates need to be more able to have the same chance of passing. Increasing the *discrimination* of an item causes the gradient of the curve to increase. Candidates below a given ability are less likely to answer correctly, whilst candidates above a given ability are more likely to answer correctly.

Figure 2 shows the latent trait model item plots. The figure shows a mix of items that resemble and do not resemble a normal ogive. Among others items 5, 9, 11, 12, 13, 15, 18, and 39 were found highly deviating from the desired normal ogive shape.

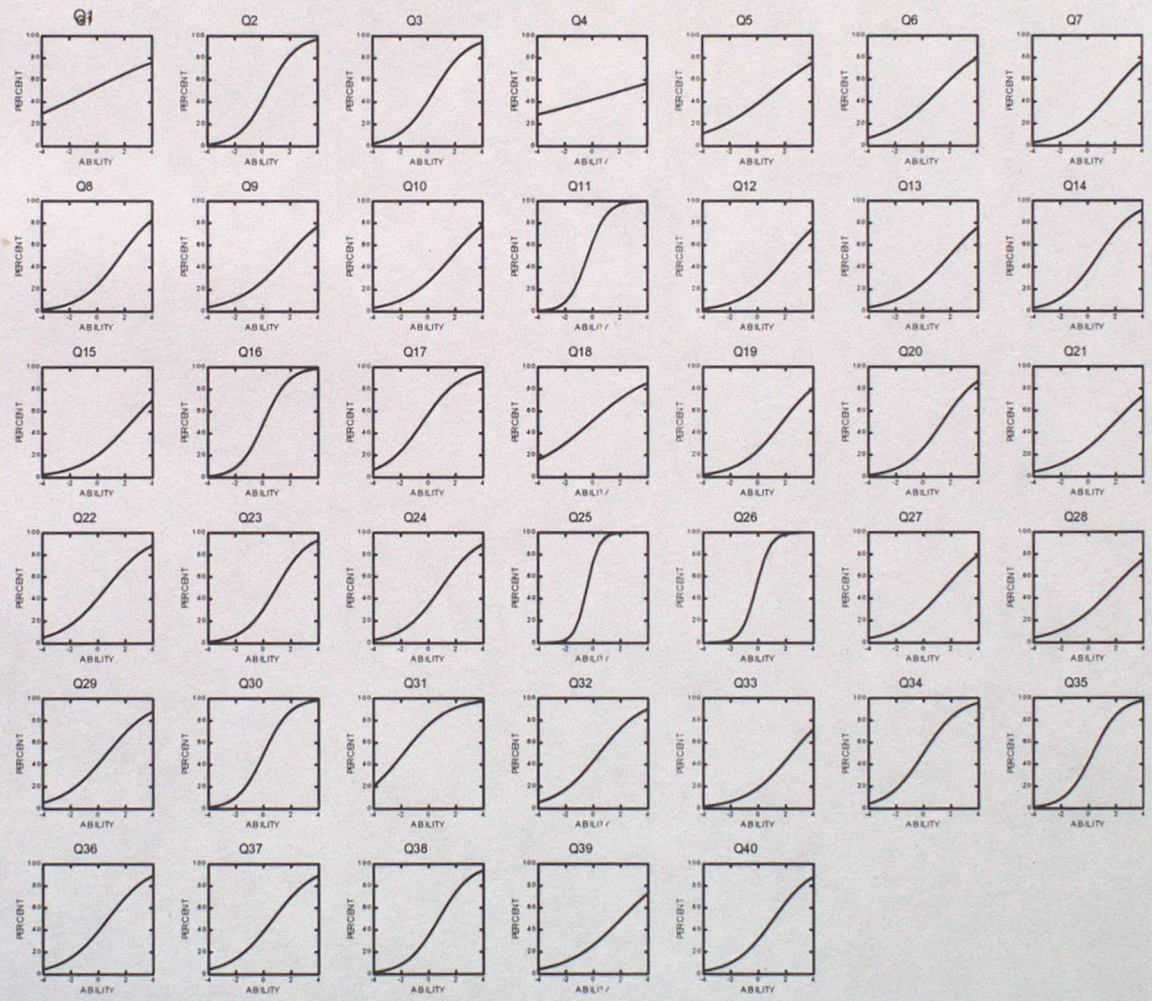


Figure 2. Latent trait model item plots for the whole Group

4.1.2 IRT parameter estimates for the whole group

The estimated parameter estimates the difficulty which is the 'b parameter' from the model. A histogram of the difficulty parameters is shown beside the difficulty estimates. The discrimination is the 'a parameter' from the model and a histogram of the discrimination parameters is shown beside the discrimination estimates. Table 3 shows the parameter estimates based on the whole group.

Table 3. IRT parameter estimates based on the whole group

Item	Difficulty	Discrimination
b1	-1.8112	0.1619
b2	0.3044	1.0062
b3	0.3754	0.8599
b4	3.6355	0.0862
b5	9.3869	0.0667
b6	1.1274	0.5124
b7	1.9236	0.5967
b8	1.2216	0.8879
b9	3.2476	0.3117
b10	2.2298	0.4724
b11	-0.3042	1.5146
b12	2.2136	0.5870
b13	3.6292	0.3020
b14	0.5655	0.8569
b15	6.7492	0.2060
b16	0.0372	1.1525
b17	-0.3940	0.7114
b18	-0.1672	0.4421
b19	1.5254	0.7160
b20	0.9547	0.9818
b21	2.0030	0.4916
b22	0.5839	0.6305
b23	0.7993	0.9583
b24	0.7072	0.8713
b25	-0.4141	1.9853
b26	-0.1334	1.8488
b27	2.1622	0.4229
b28	6.3914	0.1636
b29	0.5785	0.6487
b30	-0.0009	1.1202
b31	-2.1096	0.5367
b32	0.3677	0.6833
b33	3.1796	0.4502
b34	-0.0368	0.7997
b35	0.2368	1.0640
b36	0.6184	0.6837
b37	0.6085	0.7153
b38	0.6521	0.9075
b39	8.3356	0.1392
b40	0.8349	0.8764

4.1.3 Test information curve for the whole group

A test is a set of items; therefore, the test information at a given ability level is simply the sum of the item information at that level. The test information function (TIF) shows the distribution of the IRT ability score. The normal range is between -3 and +3. In these data due to the presence of extreme figures, the range is rather wide. When the item difficulties are clustered closely around a given value, the test information function is peaked at that point on the ability scale. The maximum amount of information depends upon the values of the discrimination parameters.

Figure 2 below shows the TIF based on the whole group and the pick is at the average value in the middle.

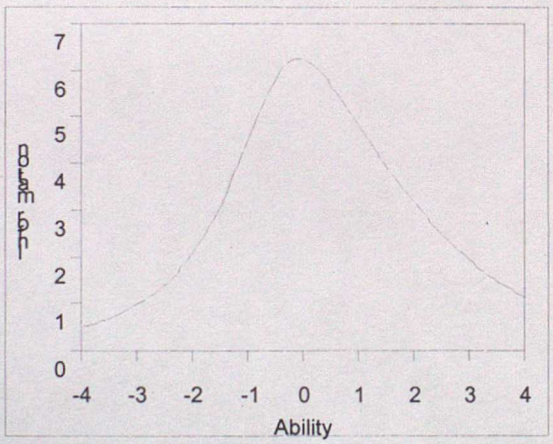


Figure 3. Test Information Function of the Whole Group

4.1.4 Item Person Dual Plots for the whole group

The information gained from item difficulty parameters in IRT models can be used to construct an increasing scale of questions, from easiest to hardest, on the same scale as the examinees. This structure gives information on which items are associated with low levels of the trait, and which are associated with high levels of the trait. Questions

are plotted to the left of the vertical dotted line, examinees on the right. In addition, a histogram of ability levels is appended to the right side of the plot.

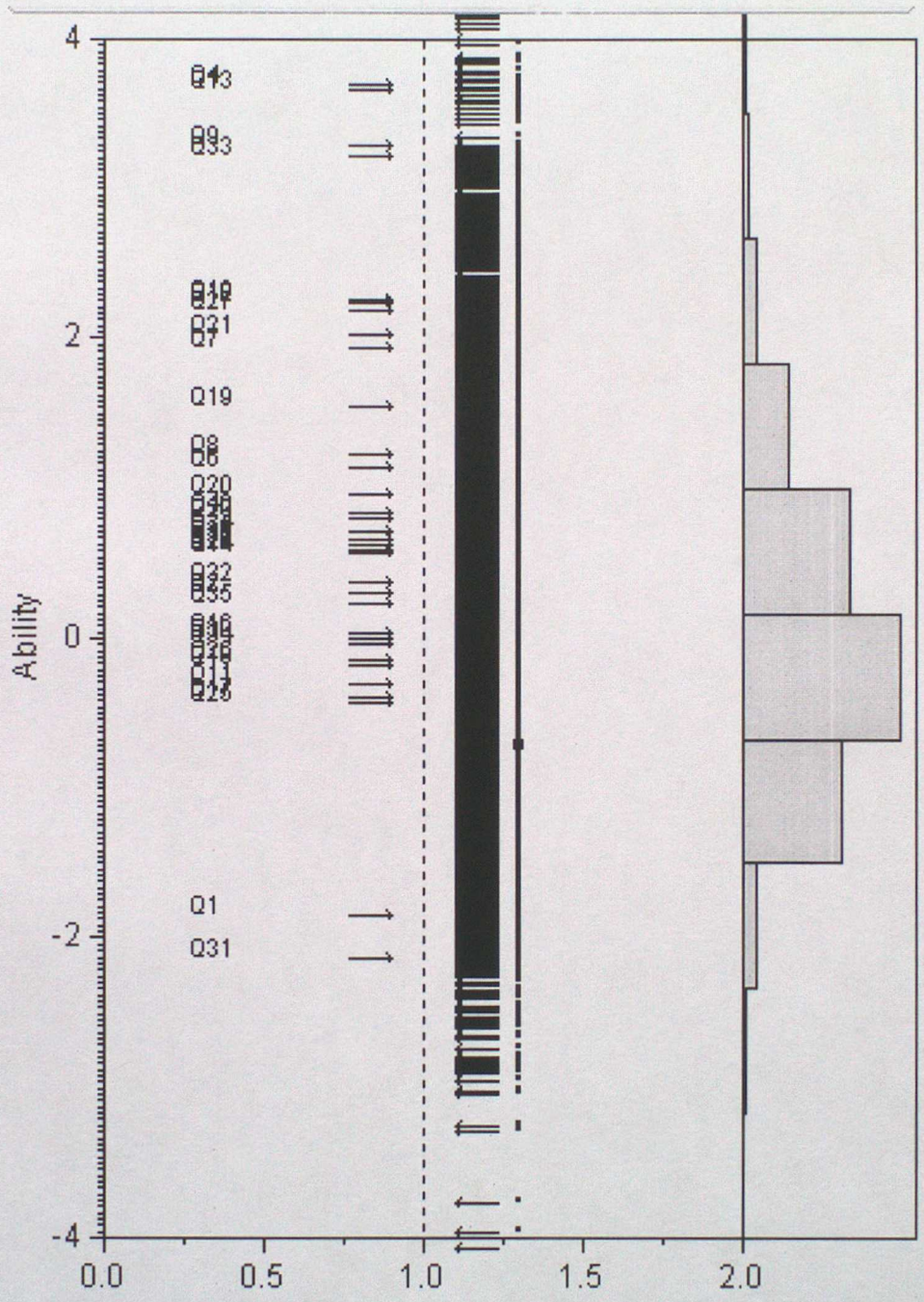


Figure 4. Item person dual plot for the whole group

The whole language group shows a wide range of abilities (Figure 4). Q13, Q9, and Q33 are rated as very difficult, with an examinee needing to be around two standard deviations above the mean in order to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels, with Q1 and Q31 appearing as easier. There are some questions that are off the displayed scale.

4.1.5 One way analysis of ability score by language

Following a two parameter logistic IRT procedure ability score was generated and appended to the main data. A one way analysis of variance using this new variable by language is computed the findings are presented below.

4.1.6 Box and whisker plots

Figure 5 below shows box and whisker plots of the IRT ability score by language groups. The mean of IRT ability scores is always 0 regardless of the language group and the box shows the confidence interval (CI). The green line which passes through each box shows the median score of the language group.

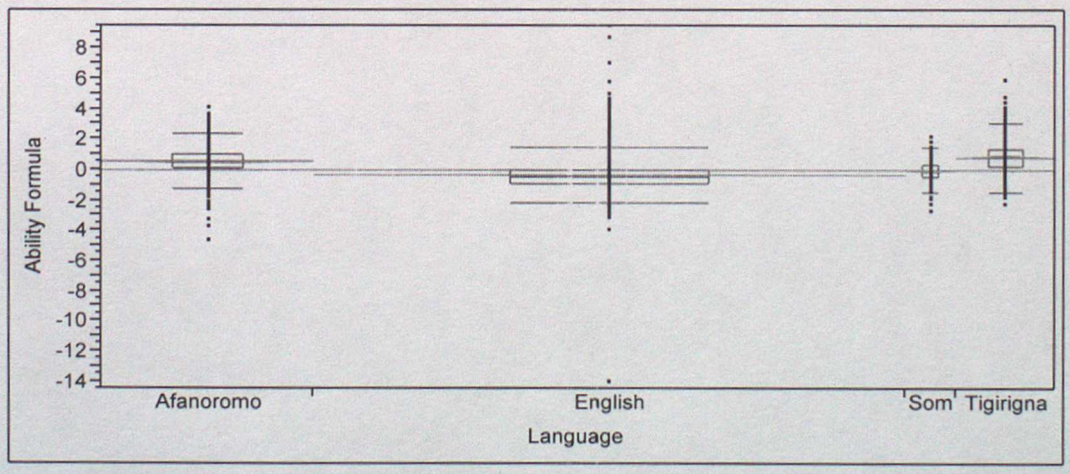


Figure 5. Box plots for ability score by language

4.1.7 Quintals of ability score by language

Table 2 below shows the quintile ability scores by the language groups. The scores vary at different ability levels across the language groups. The median score for English (-0.463) and Somali (-0.065) groups is less than the mean score. At all ability levels the Tigrigna language groups performed better than each of the other language groups.

Table 4. Quintiles of ability score by language

Language	Minimum	10%	25%	Median	75%	90%	Maximum
Afan Oromo	-4.601	-0.292	0.057	0.476	0.981	1.499	4.021
English	-14.102	-1.242	-0.873	-0.463	0.047	0.786	9.445
Somali	-2.774	-0.807	-0.446	-0.065	0.335	0.719	2.114
Tigirigna	-2.268	-0.258	0.216	0.749	1.354	2.082	5.918

A one way analysis of variance was carried out to see the existence of statistically significant differences between the scores across language groups (Table 5). The difference was statistically significant at $F_{(3, 9518)} = 797.26$, $p < .001$). The total variation explained by language was 20% ($R^2 = 0.20$).

Table 5. Analysis of variance by language

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob> F
Language	3	1912.0596	637.353	797.2631	0.0000*
Error	9518	7608.9404	0.799		
C. Total	9521	9521			

4.1.8 Recursive partitioning analysis of IRT ability score by language

The data were split at two levels and the language level variance component was found to be 0.20. In the first level the data were split into two groups where one group contained English and Somali the other group contained Afan Oromo and Tigrigna. There was no measure difference within the first group and in the second step only the

second group was split further (Figure 6).

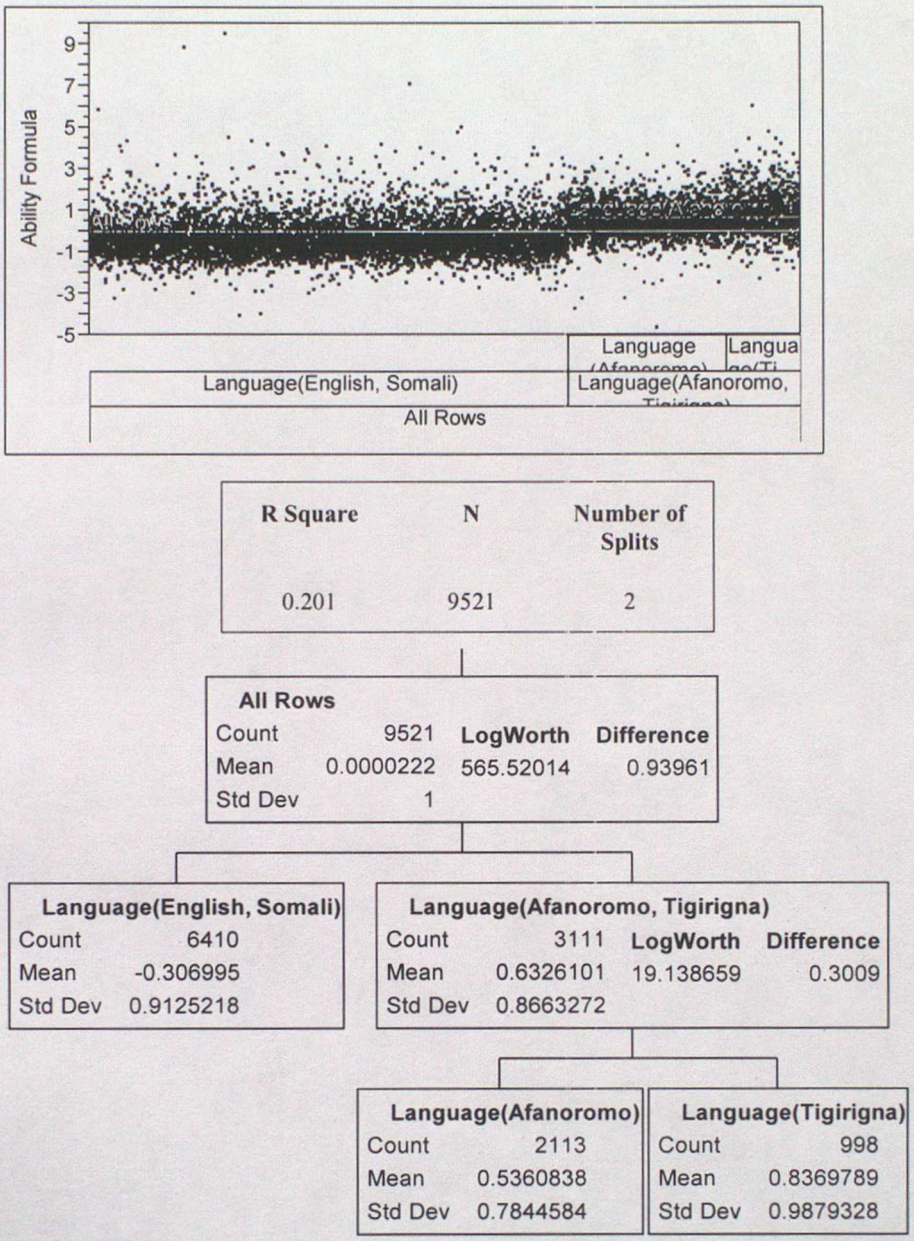


Figure 6. Recursive partitioning analysis based on IRT ability score by language

4.2 IRT results for Afan Oromo language group

In Afan Oromo language group 2,118 cases were analyzed; each containing 40 items, 5 cases were deleted by editing for missing data or for zero or perfect total scores after item editing. No item was deleted by editing for missing data or for zero or perfect total scores after item editing. Data below are based on 2,113 cases and 40 items. The

total score mean is 18.1 with a standard deviation of 4.68. This is higher than the whole group score by 2.7 points.

4.2.1 Item Characteristic Curves for the Afan Oromo language group

Figure 7 below shows the latent trait model item plots. The figure shows a mix of items that resemble and do not resemble a normal Ogive. Items 1, 4, 7, 8, 10, 13, 14, 15, 19, 21, 25, 26, and 33 were found highly deviating from the model

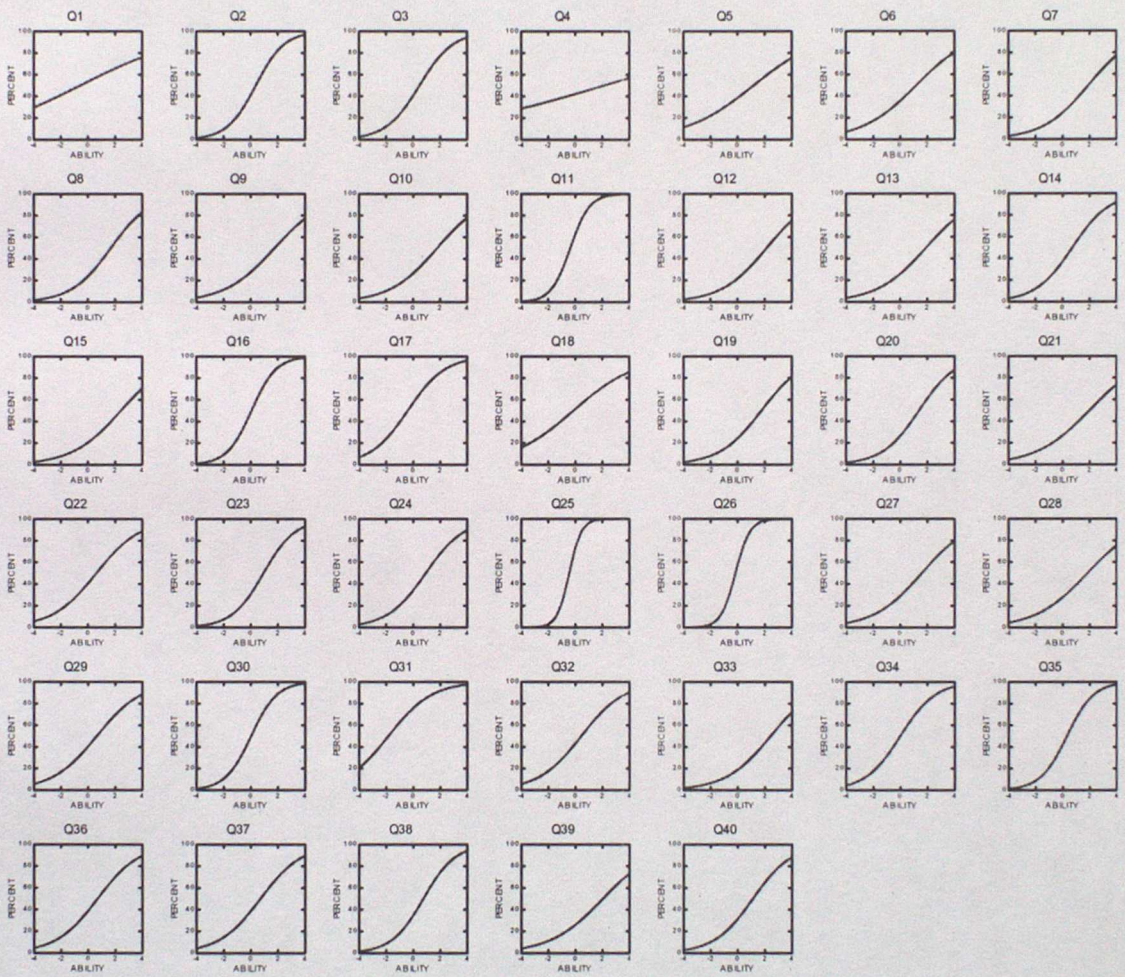


Figure 7. Latent trait model item plots for Afan Oromo language group

4.2.2 IRT parameter estimates for Afan Oromo language group

Table 6 below shows the parameter estimates based on the Afan Oromo language group.

Table 6. IRT parameter estimates for Afan Oromo language group

Item	Difficulty	Discrimination
b1	0.2160	0.2956
b2	-0.3200	1.2863
b3	-0.2578	1.4113
b4	15.9069	0.0161
b5	0.2015	0.6248
b6	-0.2709	0.4784
b7	90.3703	0.0183
b8	9.1490	0.1720
b9	2.3838	0.3077
b10	7.4005	0.2583
b11	-1.3563	0.8741
b12	-2.5785	-1.1672
b13	12.5182	0.0886
b14	0.3542	1.0269
b15	9.2075	0.1657
b16	-0.5367	1.3756
b17	-0.5908	0.6822
b18	-1.6941	0.4840
b19	1.7925	0.6990
b20	0.9730	0.9164
b21	33.9529	0.0485
b22	0.1710	0.7056
b23	0.3210	0.7737
b24	-0.0617	0.9169
b25	-1.6039	1.7655
b26	-1.6554	1.5461
b27	1.4820	0.3844
b28	-14.029	-0.0874
b29	-0.9898	0.9090
b30	-1.1601	0.6464
b31	-1.4127	0.5361
b32	-0.2355	0.5816
b33	-5.6469	-0.5377
b34	-0.8601	0.6976
b35	-0.4394	1.1337
b36	0.0453	0.8865
b37	0.4856	0.5254
b38	-0.0373	0.6556
b39	1.2546	1.0911
b40	1.4385	0.4869

4.2.3 Test information curve for Afan Oromo language group

Figure 8 below shows the TIF based on the Afan Oromo language group which is slightly positively skewed and somehow even (flatter) at the middle below the average.

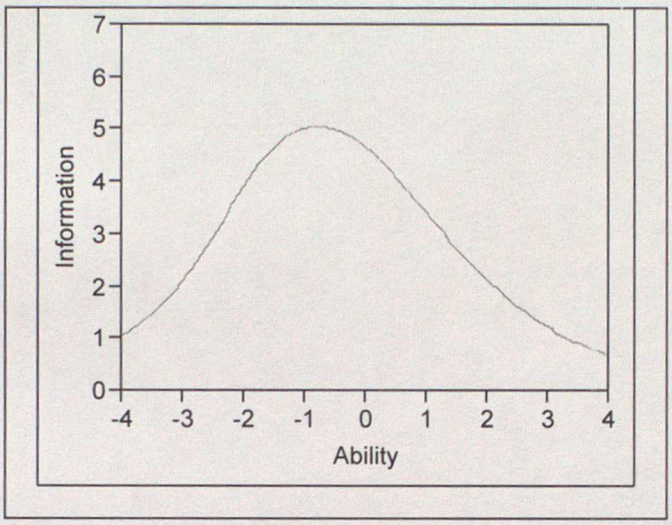


Figure 8. Test Information Function of Afan Oromo Language Group

4.2.4 Item Person Dual Plots for Afan Oromo language group

The Afan Oromo language group shows a wide range of abilities (Figure 9). Q9, Q19, Q20, Q23, Q39, and Q40 are rated as very difficult, with an examinee needing between one two standard deviations above the mean in order to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels, with Q12 appearing as easiest. There are some questions that are off the displayed scale.

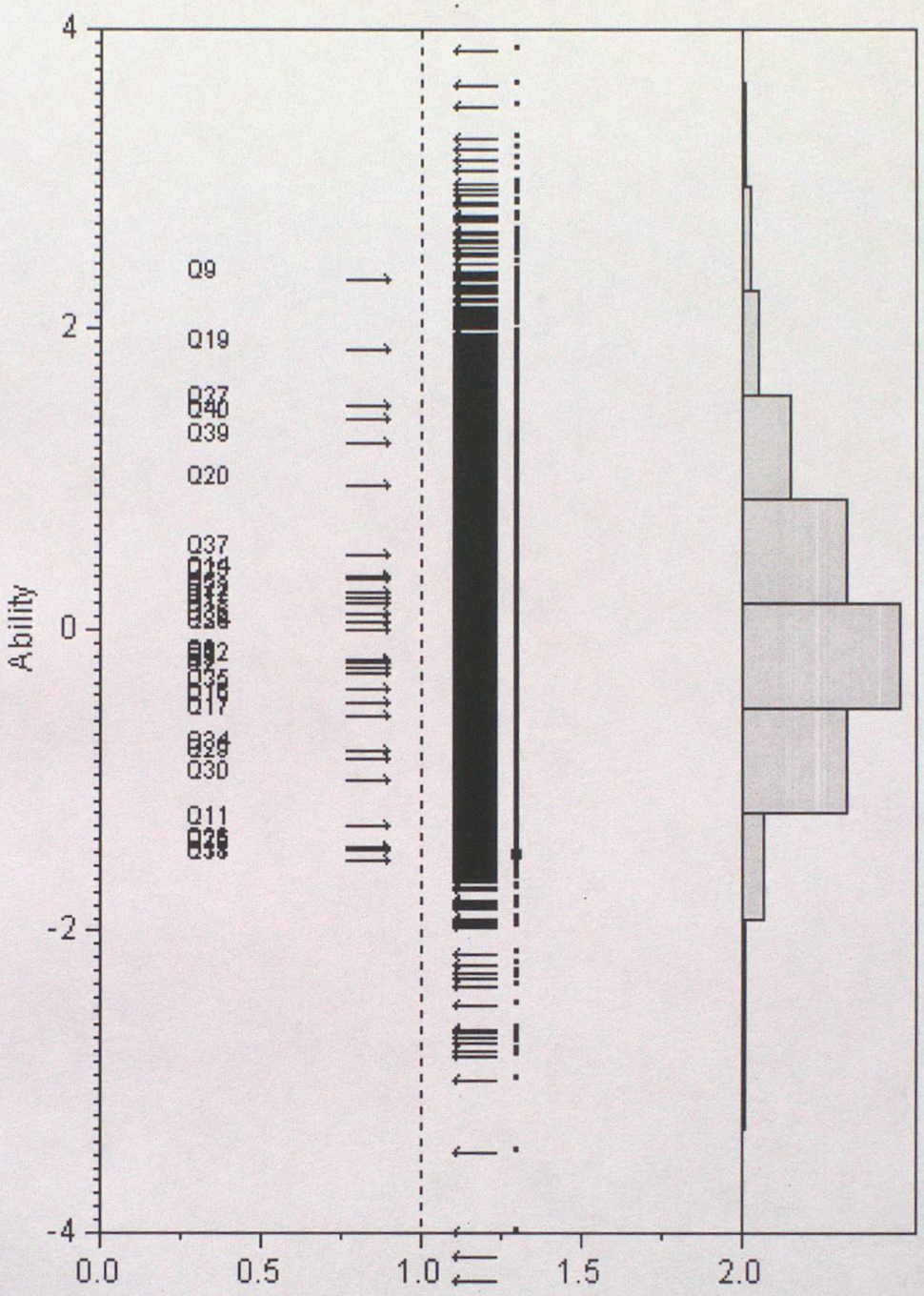


Figure 9. Item person dual plots for Afan Oromo language group

4.3 IRT results for English language group

In English language group 5,919 cases were analyzed; each containing 40 items, 5 cases were deleted by editing for missing data or for zero or, perfect total scores after item editing. No item was deleted by editing for missing data or for zero or perfect total scores after item editing. Data below are based on 5,914 cases and 40 items. The

total score mean is 13.9 with a standard deviation of 4.68. This is lower by 1.5 points than the whole group and by 4.2 from the Afan Oromo language group.

4.3.1 Item Characteristic Curves for the English language group

Figure 10 shows the latent trait model item plots. The figure shows a mix of items that resemble and do not resemble a normal Ogive. Items 1, 4, 5, 12, 13, and 15 were found somehow deviating from the model.

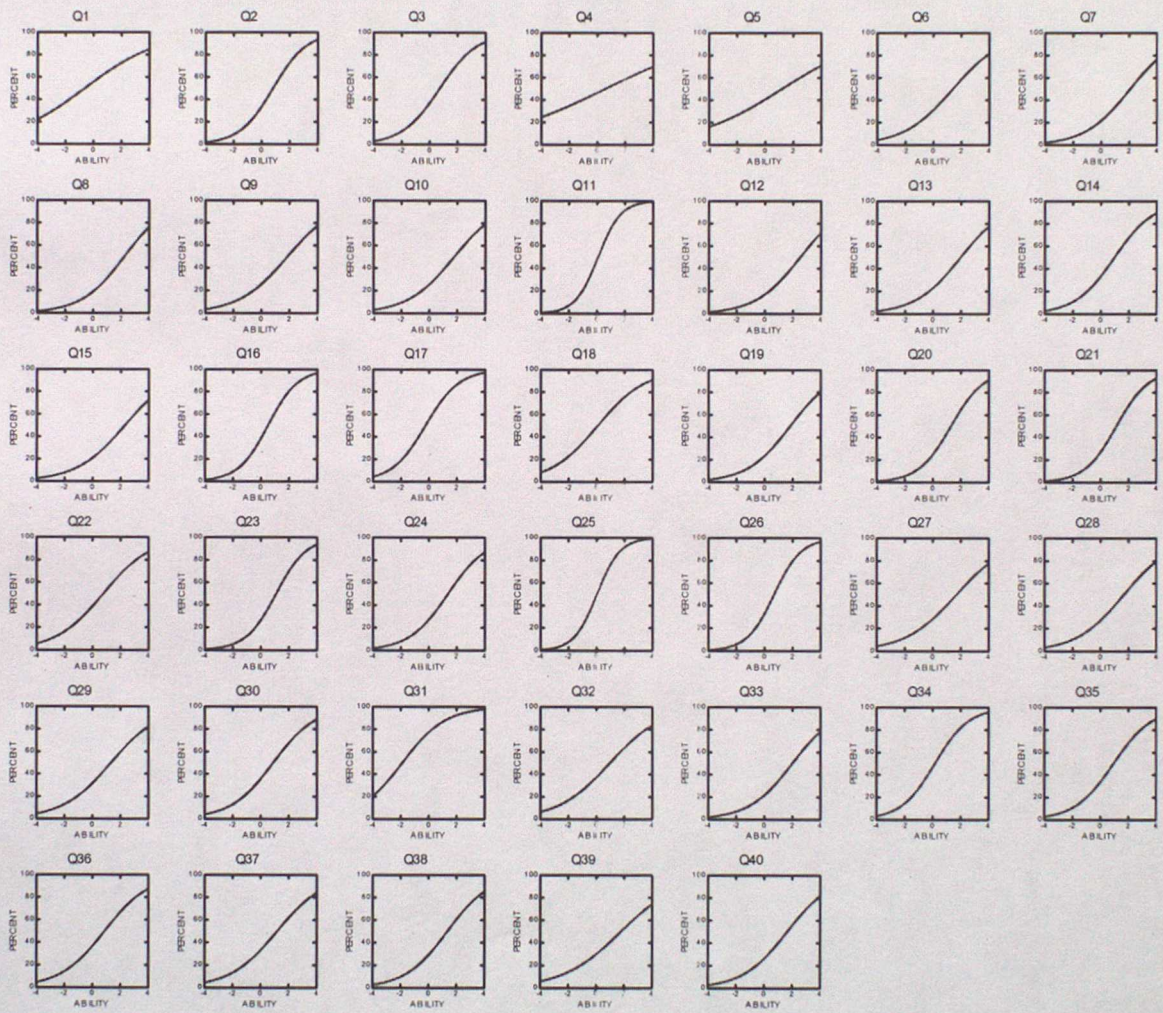


Figure 10. Latent trait model item plots for English language group

4.3.2 Test information curve for the English language group

Figure 11 below shows the TIF based on the English language group which is slightly negatively skewed and the pick is somehow above the average.

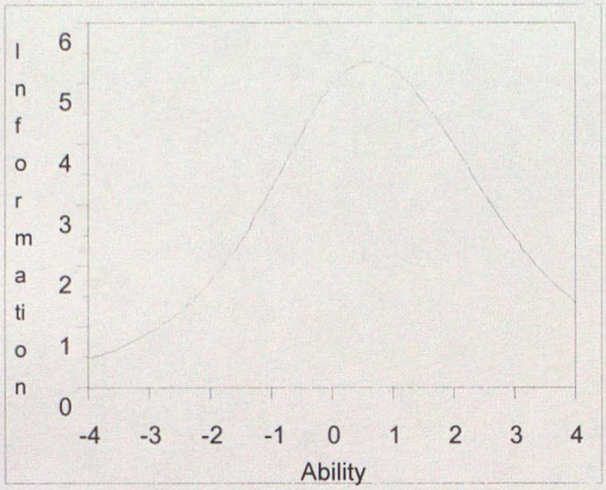


Figure 11. Test Information Function of English Language Group

4.3.3 Item Person Dual Plots for the English language group

The English language group shows a wide range of abilities (Figure 12). Q9, Q4, Q27, Q12, Q10, and Q28 are rated as very difficult, with an examinee needing between two standard deviations above the mean in order to have a 50% chance of correctly answering the question. Very few questions are distributed at lower ability levels, with Q5 appearing as easiest. There are some questions that are off the displayed scale.

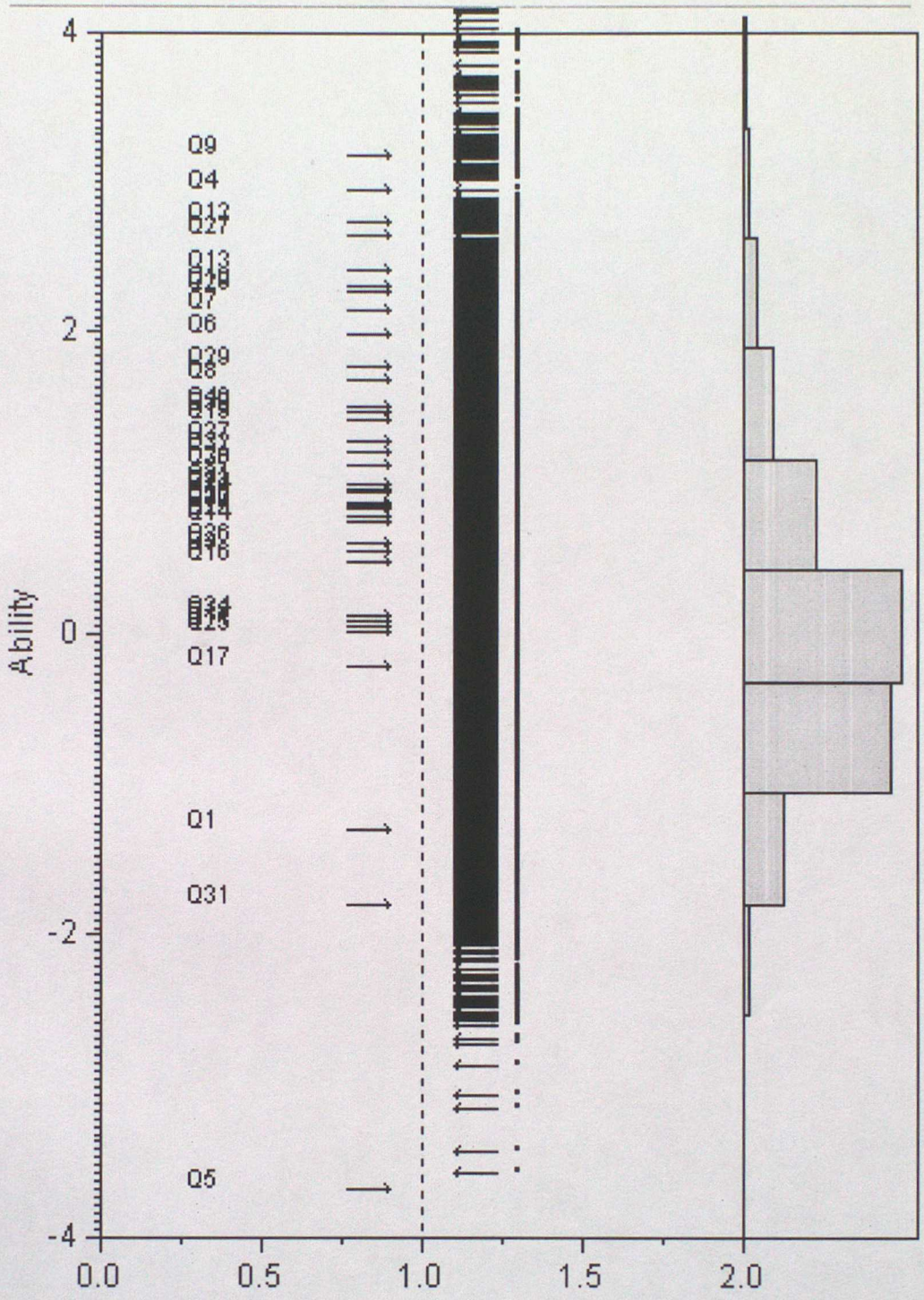


Figure 12. Item person dual plot for English language group

4.3.4 IRT parameter estimates for English language group

Table 7 below shows the parameter estimates based on the English language group.

Table 7. IRT parameter estimates for English language group

Item	Difficulty	Discrimination
b1	-1.2285	0.2881
b2	0.7473	0.8847
b3	0.5278	0.8763
b4	2.7703	0.0986
b5	-3.6193	-0.1873
b6	1.8729	0.4413
b7	1.9878	0.6344
b8	1.5754	0.9131
b9	2.9549	0.3729
b10	2.2173	0.4846
b11	0.0537	1.3577
b12	-9.5534	-0.0873
b13	2.3114	0.5188
b14	0.7209	0.8757
b15	5.4278	0.2507
b16	0.4635	1.0250
b17	-0.1962	0.8246
b18	0.0794	0.5676
b19	1.3499	0.8746
b20	1.0851	1.0349
b21	0.9333	1.0519
b22	0.8520	0.5609
b23	0.9342	1.1060
b24	0.9337	0.9947
b25	0.0208	1.3432
b26	0.5706	1.1464
b27	2.4792	0.4035
b28	2.1288	0.4617
b29	1.6779	0.4715
b30	0.7873	0.7346
b31	-1.6740	0.7120
b32	1.1334	0.5064
b33	1.4127	0.9197
b34	0.1174	0.8961
b35	0.8203	0.7901
b36	0.8186	0.6673
b37	1.2135	0.5943
b38	1.0696	0.8057
b39	59.7281	0.0173
b40	1.4476	0.7142

4.4 IRT results for Somali language group

In Somali language group 498 cases were analyzed; each containing 40 items 1 case was deleted by editing for missing data or for zero or perfect total scores after item

editing. No item was deleted by editing for missing data or for zero or perfect total scores after item editing. The total score mean is 15.4 with a standard deviation of 4.08. This is equal to the whole group but lower than Afan Oromo language group by 2.7 points.

4.4.1 Item Characteristic Curves for the Somali language group

Figure 13 shows the latent trait model item plots. The figure shows a mix of items that resemble and do not resemble a normal ogive. Items 1, 2, 3, 4, 7, 9, 11, 15, 21, 24, 27, 28, 33, 34, and 36 did not work as expected and deviated highly from the model.

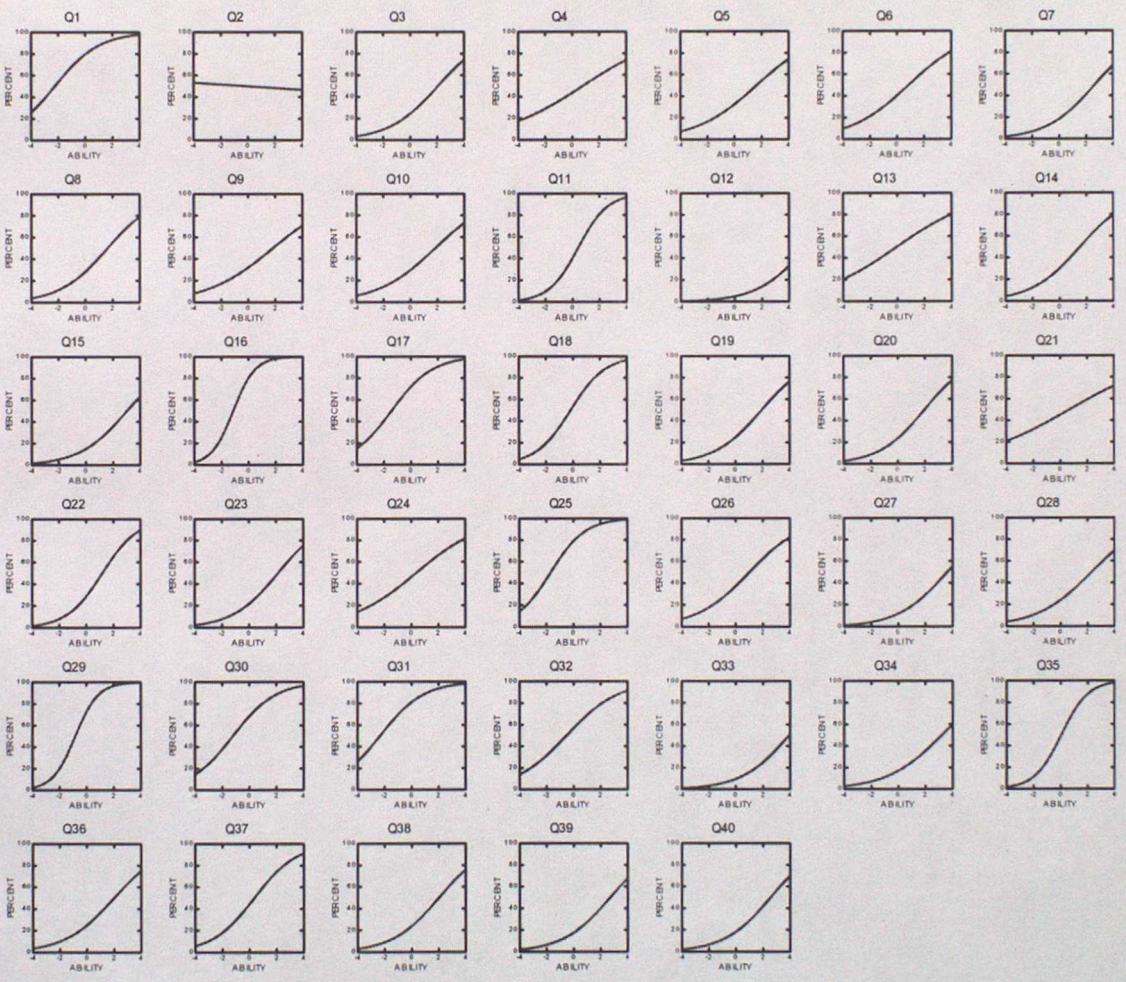


Figure 13. Latent trait model item plots for Somali language group

4.4.2 Test information curve for the Somali language group

Figure 14 below shows the TIF based on the Somali language group which is positively skewed and the pick is below the average.

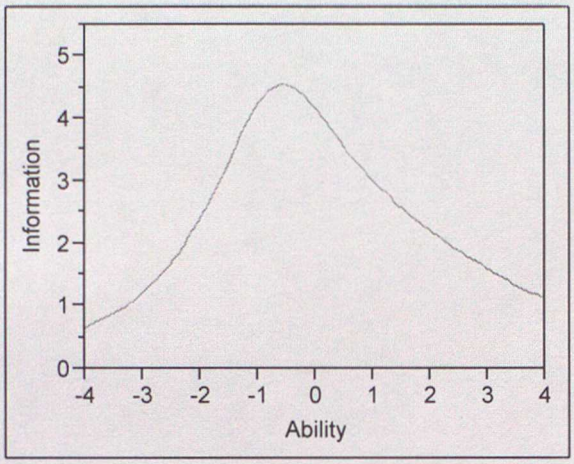


Figure 14. Test Information Function of Somali Language Group

4.4.3 IRT parameter estimates for Somali language group

The Somali language group shows a wide range of abilities (Figure 15). Q15, Q4, Q39, Q33, Q36, and Q7 are rated as very difficult, with an examinee needing around three standard deviations above the mean in order to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels, with Q1 appearing as easiest. There are some questions that are off the displayed scale.

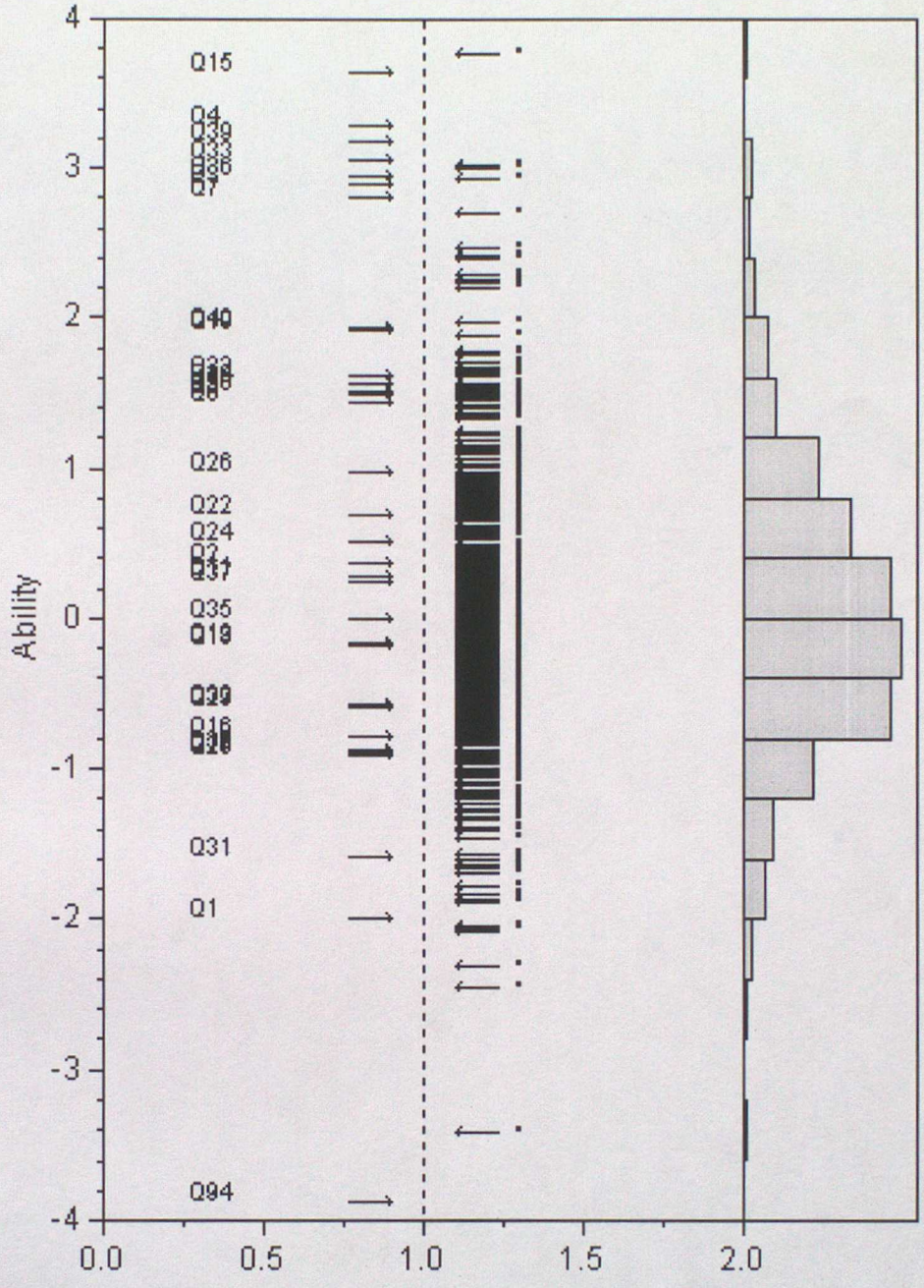


Figure 15. Item person dual plot for Somali language group

4.4.4 IRT parameter estimates for English language group

Table 8 shows the parameter estimates based on the Somali language group.

Table 8. IRT parameter estimates for Somali language group

Item	Difficulty	Discrimination
b1	-1.9988	0.6607
b2	0.3704	-0.1261
b3	2.8785	0.3889
b4	3.2393	0.1229
b5	9.5096	0.0905
b6	1.4251	0.3105
b7	2.7877	0.5227
b8	1.4990	0.6086
b9	-3.8990	-0.2460
b10	15.4432	0.0640
b11	0.2809	1.0165
b12	-0.9218	0.0353
b13	-0.1646	0.3217
b14	1.5670	0.5330
b15	3.6326	0.4601
b16	-0.7786	1.8467
b17	-0.8903	1.0288
b18	-0.1806	0.7761
b19	1.9048	0.5636
b20	1.5764	0.6999
b21	-7.8940	-0.0422
b22	0.6989	1.0389
b23	1.5985	0.7682
b24	0.5172	0.3395
b25	-0.9036	1.2834
b26	0.9730	0.5544
b27	42.8967	0.0455
b28	30.4980	0.0406
b29	-0.5894	1.7769
b30	-0.8729	0.8681
b31	-1.5821	0.9004
b32	-0.5687	0.5608
b33	3.0221	0.7390
b34	-3.8760	-0.4579
b35	0.0024	1.1121
b36	2.9148	0.3725
b37	0.2540	0.7049
b38	1.5000	0.7648
b39	3.1313	0.4758
b40	1.9276	0.7580

4.5 IRT results for Tigrigna language group

In Tigrigna language group 998 cases were processed, each containing 40 items. Neither a case nor an item was deleted by editing for missing data or for zero or

perfect total scores after item editing. Data below are based on 998 cases and 40 Items. The total score mean is 19.2 with a standard deviation of 5.57. This is the highest score of all language groups, the widest difference being with English (5.3 points).

4.5.1 Item Characteristic Curves for the Tigrigna language group

Figure 16 shows the latent trait model item plots. The figure shows a mix of items that resemble and do not resemble a normal Ogive. Items 1, 11, 12, 15, 18, 23, and 26 did not work as expected and deviated from the model a lot.

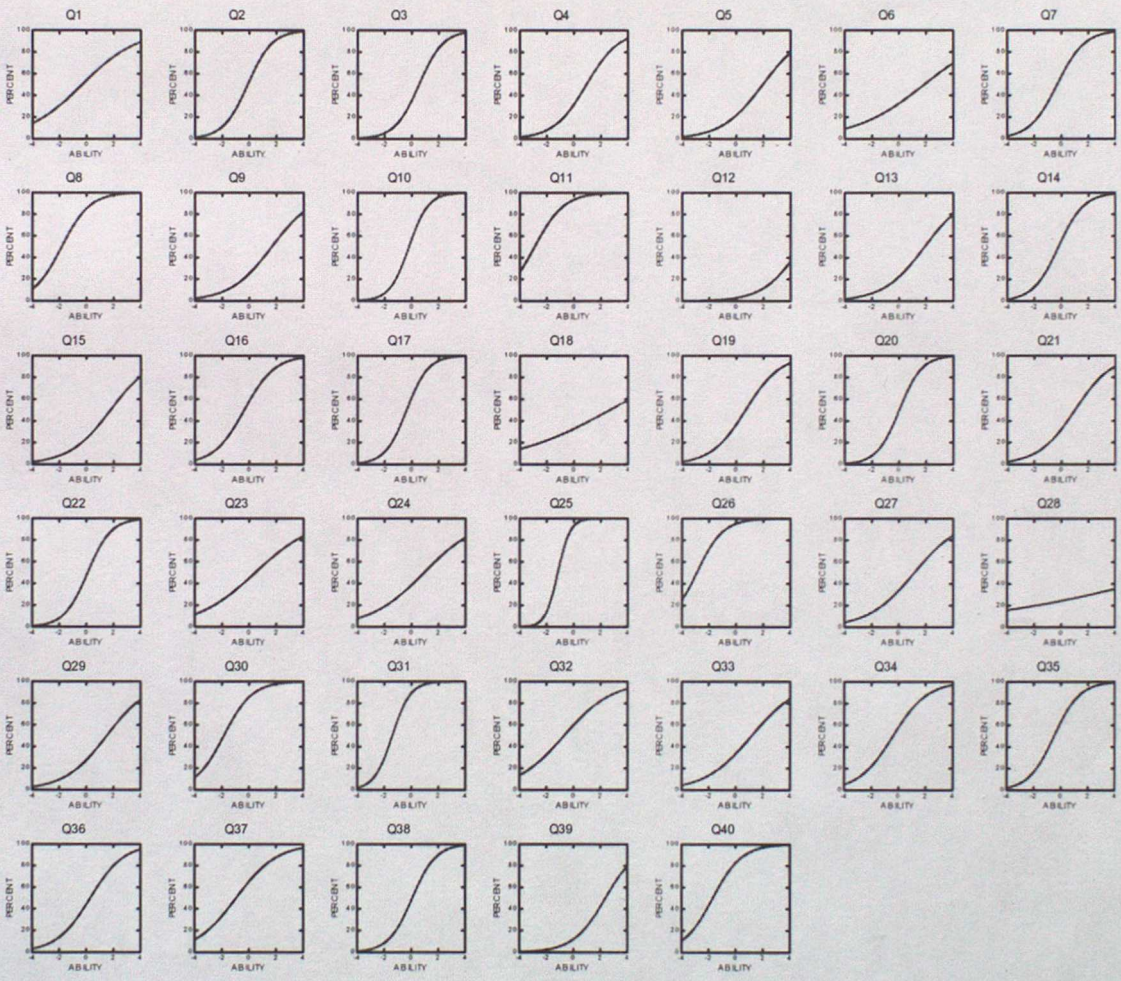


Figure 16. Latent trait model item plots for Tigrigna

4.5.2 Test information curve for the Tigrigna language group

Figure 8 below shows the TIF based on the Tigrigna language group which is positively skewed the pick is below the average by one standard deviation.

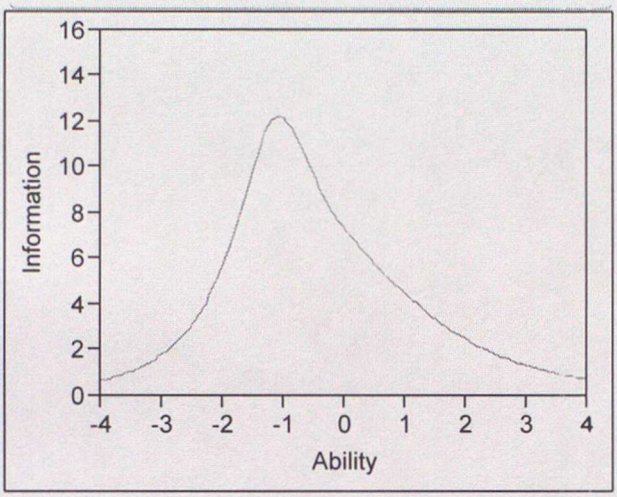


Figure 17. Test Information Function of Tigrigna Language Group

4.5.3 Item Person Dual Plots for the Tigrigna language group

The Tigrigna language group shows a wide range of abilities (Figure 18). Q9, Q15, and Q29 are rated as very difficult, with an examinee needing around two and half standard deviations above the mean in order to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels, with Q12 appearing as easiest. There are some questions that are off the displayed scale.

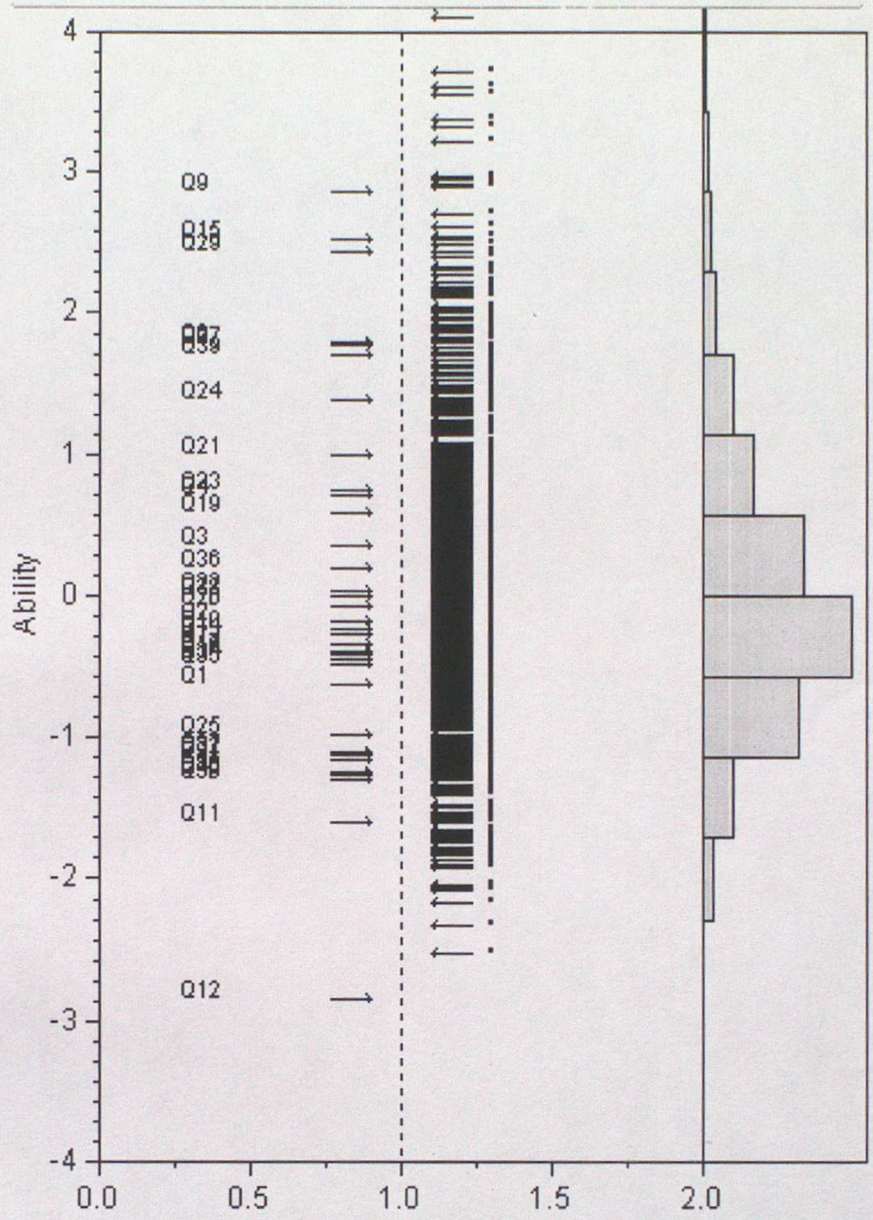


Figure 18. Item person dual plot for Tigrigna language group

4.5.4 IRT parameter estimates for Tigrigna language group

Table 9 below shows the parameter estimates based on the Tigrigna language group.

Table 9. IRT parameter estimates for Tigrigna language group

Item	Difficulty	Discrimination
b1	-0.6201	0.4604
b2	-0.1854	1.1603
b3	0.3636	1.2462
b4	0.7068	0.9612
b5	4.6202	0.2981
b6	1.7790	0.3951
b7	-0.3384	1.0469
b8	-1.2424	1.4561
b9	2.7889	0.4198
b10	-0.2371	1.5021
b11	-1.6182	1.8117
b12	-14.402	-0.1766
b13	4.4009	0.3017
b14	-0.2618	1.1315
b15	2.4887	0.4988
b16	-0.4082	0.9430
b17	-0.3918	1.3640
b18	5.9094	0.1261
b19	0.5912	0.7938
b20	-0.0643	1.3655
b21	1.0002	0.7113
b22	0.0290	1.2764
b23	0.7429	0.3646
b24	1.4109	0.4113
b25	-0.9744	3.0615
b26	-1.2391	3.4407
b27	1.7597	0.4210
b28	-12.058	-0.0973
b29	2.4279	0.4329
b30	-1.3026	1.3875
b31	-1.1549	1.9948
b32	-1.1301	0.4913
b33	1.7046	0.4434
b34	-0.4451	0.8718
b35	-0.4823	1.1577
b36	0.1982	0.8561
b37	-1.1063	0.6882
b38	0.0095	1.2238
b39	1.7150	1.1160
b40	-1.2725	1.2543

4.6 Findings based on Classical Test Theory

Classical Test Theory does not provide information about how testers at different ability levels perform on the item. Item response theory (IRT) was developed to overcome the preceding shortcoming. In IRT a test is unbiased if all testers having the same skill level have an equal probability of getting the item correct regardless of group membership.

4.7 CTT results for the whole group

For the whole group 9,533 cases were processed. There was no missing data and the analysis was based on 9,533 complete cases for 40 data items. A total of 10 items namely Item 1, 4, 5, 9, 12, 13, 15, 28, 33 and 39 were found to be problem items (Table 10). The discrimination indexes of these items were found below 0.2 and in case of Item 12 it is even negative. These items should have been substantially modified or discarded altogether. Items 9, 12, 13, 15, 28, 33 and 39 were found very difficult to the group and were answered only by less than 30% of the test-takers in each case.

Table 10.Item Analysis Results for the Whole Group

Number Item	Key	Item Correct	Diff.	Disc. Index	# Correct in High Grp		# Correct in Low Grp		Point Biser.	Adj. Pt Bis
Item 01	(2)#	5451	0.57	0.18	1873	(0.64)	1422	(0.46)	0.15	0.06
Item 02	(4)	4076	0.43	0.45	2002	(0.68)	726	(0.23)	0.39	0.31
Item 03	(3)	4032	0.42	0.40	1897	(0.64)	756	(0.24)	0.36	0.27
Item 04	(1)#	4023	0.42	0.11	1401	(0.48)	1141	(0.37)	0.14	0.05
Item 05	(2)#	3320	0.35	0.11	1188	(0.40)	913	(0.29)	0.11	0.02
Item 06	(3)	3455	0.36	0.27	1515	(0.51)	760	(0.25)	0.25	0.16
Item 07	(1)	2377	0.25	0.23	1125	(0.38)	459	(0.15)	0.26	0.18
Item 08	(4)	2575	0.27	0.33	1352	(0.46)	399	(0.13)	0.33	0.26
Item 09	(2)#	2558	0.27	0.18	1082	(0.37)	578	(0.19)	0.19	0.11
Item 10	(1)	2512	0.26	0.20	1129	(0.38)	565	(0.18)	0.24	0.16
Item 11	(4)	5510	0.58	0.53	2482	(0.84)	955	(0.31)	0.45	0.37
Item 12	(1)#	2279	0.24	-0.05	576	(0.20)	761	(0.25)	-0.05	-0.12
Item 13	(1)#	2405	0.25	0.17	1023	(0.35)	553	(0.18)	0.19	0.11
Item 14	(4)	3697	0.39	0.37	1777	(0.60)	713	(0.23)	0.35	0.27
Item 15	(3)#	1908	0.20	0.10	775	(0.26)	492	(0.16)	0.14	0.07
Item 16	(2)	4609	0.48	0.49	2199	(0.75)	803	(0.26)	0.41	0.33
Item 17	(1)	5347	0.56	0.36	2210	(0.75)	1201	(0.39)	0.32	0.24
Item 18	(2)	4917	0.52	0.27	1886	(0.64)	1135	(0.37)	0.23	0.14
Item 19	(4)	2507	0.26	0.27	1274	(0.43)	490	(0.16)	0.31	0.23
Item 20	(3)	2861	0.30	0.35	1534	(0.52)	514	(0.17)	0.38	0.30
Item 21	(1)	2641	0.28	0.23	1183	(0.40)	535	(0.17)	0.27	0.19
Item 22	(2)	3918	0.41	0.31	1720	(0.58)	833	(0.27)	0.30	0.21
Item 23	(4)	3166	0.33	0.39	1635	(0.55)	523	(0.17)	0.37	0.29
Item 24	(4)	3435	0.36	0.39	1693	(0.57)	557	(0.18)	0.34	0.26
Item 25	(2)	5960	0.63	0.57	2686	(0.91)	1065	(0.34)	0.47	0.40
Item 26	(2)	5054	0.53	0.58	2451	(0.83)	780	(0.25)	0.47	0.40
Item 27	(4)	2760	0.29	0.20	1189	(0.40)	627	(0.20)	0.22	0.14
Item 28	(3)#	2482	0.26	0.11	937	(0.32)	646	(0.21)	0.13	0.05
Item 29	(2)	3904	0.41	0.33	1707	(0.58)	776	(0.25)	0.28	0.20
Item 30	(3)	4700	0.49	0.46	2160	(0.73)	842	(0.27)	0.39	0.31
Item 31	(2)	7111	0.75	0.23	2520	(0.86)	1932	(0.62)	0.24	0.17
Item 32	(4)	4176	0.44	0.34	1803	(0.61)	839	(0.27)	0.30	0.21
Item 33	(4)#	1887	0.20	0.16	863	(0.29)	423	(0.14)	0.22	0.15
Item 34	(1)	4787	0.50	0.38	2069	(0.70)	999	(0.32)	0.33	0.24
Item 35	(2)	4189	0.44	0.45	2032	(0.69)	737	(0.24)	0.39	0.31
Item 36	(3)	3805	0.40	0.35	1749	(0.59)	769	(0.25)	0.31	0.23
Item 37	(4)	3782	0.40	0.36	1716	(0.58)	692	(0.22)	0.31	0.22
Item 38	(3)	3490	0.37	0.38	1711	(0.58)	630	(0.20)	0.36	0.28
Item 39	(2)#	2277	0.24	0.09	889	(0.30)	667	(0.22)	0.14	0.06
Item 40	(1)	3211	0.34	0.36	1600	(0.54)	573	(0.18)	0.34	0.26

4.8 CTT results for Afan Oromo language group

In Afan Oromo language group 2,118 cases were processed. There is no missing data and analysis was based on 2118 complete cases for 40 data items. A total of 12 items namely Item 4, 7, 8, 9, 10, 12, 13, 15, 21, 25, 28, and 33 were found to be problem items (Table 11). The discrimination indexes of these items were found below 0.2 and in case of Item 12 and 33 it is even negative. These items should have been substantially modified or discarded altogether. Items 7, 8, 12, 13, 15, 21, 28, 33, were

found very difficult to the group and answered only by less than 30% of the test-takers in each case.

Table 11.Item Analysis Results for Afan Oromo Language Group

Number Item	Key	Item Correct		Disc. Index	# Correct in High Grp		# Correct in Low Grp		Point Biser.	Adj. Pt Bis
Item 01	(2)	982	0.48	0.21	403	(0.60)	247	(0.38)	0.21	0.10
Item 02	(4)	1172	0.58	0.48	558	(0.83)	227	(0.35)	0.40	0.31
Item 03	(3)	1146	0.56	0.48	538	(0.80)	203	(0.31)	0.42	0.32
Item 04	(1)#	889	0.44	0.12	338	(0.50)	247	(0.38)	0.13	0.02
Item 05	(2)	948	0.47	0.30	415	(0.61)	203	(0.31)	0.28	0.18
Item 06	(3)	1075	0.53	0.30	465	(0.69)	251	(0.39)	0.25	0.15
Item 07	(1)#	328	0.16	0.07	135	(0.20)	83	(0.13)	0.09	0.01
Item 08	(4)#	350	0.17	0.10	153	(0.23)	85	(0.13)	0.12	0.03
Item 09	(2)#	661	0.32	0.19	280	(0.41)	145	(0.22)	0.18	0.08
Item 10	(1)#	264	0.13	0.08	123	(0.18)	66	(0.10)	0.14	0.06
Item 11	(4)	1512	0.74	0.28	600	(0.89)	392	(0.60)	0.29	0.20
Item 12	(1)#	135	0.07	-0.06	20	(0.03)	60	(0.09)	-0.10	-0.16
Item 13	(1)#	505	0.25	0.15	218	(0.32)	115	(0.18)	0.14	0.05
Item 14	(4)	841	0.41	0.40	426	(0.63)	152	(0.23)	0.35	0.25
Item 15	(3)#	364	0.18	0.10	153	(0.23)	83	(0.13)	0.13	0.04
Item 16	(2)	1298	0.64	0.45	571	(0.84)	255	(0.39)	0.40	0.31
Item 17	(1)	1199	0.59	0.31	501	(0.74)	281	(0.43)	0.30	0.19
Item 18	(2)	1397	0.69	0.25	536	(0.79)	354	(0.55)	0.23	0.13
Item 19	(4)	471	0.23	0.24	255	(0.38)	92	(0.14)	0.27	0.18
Item 20	(3)	617	0.30	0.33	341	(0.50)	111	(0.17)	0.34	0.25
Item 21	(1)#	329	0.16	0.07	145	(0.21)	95	(0.15)	0.10	0.02
Item 22	(2)	950	0.47	0.32	432	(0.64)	208	(0.32)	0.31	0.20
Item 23	(4)	890	0.44	0.34	420	(0.62)	183	(0.28)	0.32	0.22
Item 24	(4)	1030	0.51	0.39	461	(0.68)	192	(0.30)	0.34	0.23
Item 25	(2)#	1828	0.90	0.19	660	(0.98)	507	(0.78)	0.32	0.26
Item 26	(2)	1804	0.89	0.21	650	(0.96)	489	(0.75)	0.32	0.25
Item 27	(4)	736	0.36	0.21	314	(0.46)	166	(0.26)	0.22	0.12
Item 28	(3)#	464	0.23	0.06	169	(0.25)	123	(0.19)	0.06	-0.03
Item 29	(2)	1400	0.69	0.34	567	(0.84)	324	(0.50)	0.31	0.22
Item 30	(3)	1357	0.67	0.28	529	(0.78)	328	(0.51)	0.28	0.18
Item 31	(2)	1367	0.67	0.24	527	(0.78)	348	(0.54)	0.25	0.15
Item 32	(4)	1076	0.53	0.31	454	(0.67)	233	(0.36)	0.27	0.17
Item 33	(4)#	102	0.05	-0.01	30	(0.04)	35	(0.05)	-0.01	-0.06
Item 34	(1)	1288	0.63	0.30	525	(0.78)	312	(0.48)	0.29	0.19
Item 35	(2)	1217	0.60	0.44	551	(0.82)	245	(0.38)	0.38	0.28
Item 36	(3)	986	0.48	0.37	459	(0.68)	199	(0.31)	0.34	0.23
Item 37	(4)	886	0.43	0.26	377	(0.56)	191	(0.29)	0.25	0.14
Item 38	(3)	1020	0.50	0.31	441	(0.65)	221	(0.34)	0.29	0.18
Item 39	(2)	460	0.23	0.32	291	(0.43)	71	(0.11)	0.36	0.28
Item 40	(1)	679	0.33	0.26	321	(0.47)	138	(0.21)	0.25	0.15

4.9 CTT results for English language group

In English language group 5,919 cases were analyzed. There were no missing data. Data below are based on 5919 complete cases for 40 data items. A total of 10 items namely Item 1, 4, 5, 9, 12, 13, 27, 28, and 39 were found to be problem items (Table 12). Except Item 1, the discrimination indexes of these items were found below 0.2

and unlike to the previous groups there are no negatively discriminating items. Items 9, 12, 13, 15, 27, 28, and 39, were found very difficult to the group and were answered only by less than 30% of the test-takers in each case.

Table 12.Item analysis result for English language group

Number Item	Key	Item Correct	Item Diff.	Disc. Index	# Correct in High Grp	# Correct in Low Grp	Point Biser.	Adj. Pt Bis
Item 01	(2)#	3521	0.59	0.24	1233 (0.69)	923 (0.45)	0.17	0.08
Item 02	(4)	2122	0.35	0.35	1012 (0.56)	432 (0.21)	0.36	0.27
Item 03	(3)	2374	0.40	0.36	1085 (0.60)	492 (0.24)	0.34	0.25
Item 04	(1)#	2594	0.43	0.15	901 (0.50)	723 (0.35)	0.15	0.05
Item 05	(2)#	2023	0.34	0.05	627 (0.35)	612 (0.30)	0.05	-0.04
Item 06	(3)	1856	0.31	0.20	764 (0.42)	461 (0.22)	0.23	0.14
Item 07	(1)	1392	0.23	0.23	670 (0.37)	289 (0.14)	0.26	0.18
Item 08	(4)	1280	0.21	0.27	686 (0.38)	230 (0.11)	0.32	0.24
Item 09	(2)#	1521	0.25	0.18	636 (0.35)	359 (0.17)	0.20	0.12
Item 10	(1)	1567	0.26	0.20	671 (0.37)	355 (0.17)	0.23	0.14
Item 11	(4)	2886	0.48	0.50	1350 (0.75)	515 (0.25)	0.42	0.33
Item 12	(1)#	1817	0.30	0.09	594 (0.33)	498 (0.24)	0.07	-0.02
Item 13	(1)#	1436	0.24	0.18	626 (0.35)	347 (0.17)	0.24	0.16
Item 14	(4)	2158	0.36	0.35	1017 (0.57)	440 (0.21)	0.33	0.24
Item 15	(3)#	1237	0.21	0.10	472 (0.26)	325 (0.16)	0.14	0.07
Item 16	(2)	2371	0.40	0.38	1120 (0.62)	497 (0.24)	0.37	0.28
Item 17	(1)	3218	0.54	0.37	1326 (0.74)	762 (0.37)	0.32	0.23
Item 18	(2)	2936	0.49	0.32	1170 (0.65)	687 (0.33)	0.27	0.17
Item 19	(4)	1524	0.25	0.28	786 (0.44)	317 (0.15)	0.33	0.25
Item 20	(3)	1619	0.27	0.31	841 (0.47)	321 (0.16)	0.36	0.28
Item 21	(1)	1774	0.30	0.35	909 (0.51)	311 (0.15)	0.38	0.30
Item 22	(2)	2326	0.39	0.27	975 (0.54)	562 (0.27)	0.27	0.18
Item 23	(4)	1729	0.29	0.35	908 (0.51)	311 (0.15)	0.38	0.30
Item 24	(4)	1822	0.30	0.36	912 (0.51)	309 (0.15)	0.34	0.26
Item 25	(2)	2942	0.49	0.47	1365 (0.76)	601 (0.29)	0.41	0.32
Item 26	(2)	2165	0.36	0.40	1085 (0.60)	417 (0.20)	0.39	0.31
Item 27	(4)#	1641	0.27	0.17	671 (0.37)	408 (0.20)	0.21	0.13
Item 28	(3)#	1669	0.28	0.19	690 (0.38)	402 (0.20)	0.21	0.13
Item 29	(2)	1904	0.32	0.20	778 (0.43)	469 (0.23)	0.24	0.15
Item 30	(3)	2209	0.37	0.31	973 (0.54)	476 (0.23)	0.31	0.22
Item 31	(2)	4527	0.75	0.27	1586 (0.88)	1258 (0.61)	0.26	0.18
Item 32	(4)	2191	0.37	0.25	900 (0.50)	510 (0.25)	0.25	0.15
Item 33	(4)	1414	0.24	0.27	720 (0.40)	267 (0.13)	0.33	0.25
Item 34	(1)	2851	0.48	0.38	1241 (0.69)	629 (0.31)	0.34	0.25
Item 35	(2)	2126	0.35	0.32	975 (0.54)	448 (0.22)	0.32	0.24
Item 36	(3)	2244	0.37	0.32	1003 (0.56)	479 (0.23)	0.29	0.21
Item 37	(4)	2008	0.33	0.30	886 (0.49)	404 (0.20)	0.27	0.18
Item 38	(3)	1866	0.31	0.32	897 (0.50)	378 (0.18)	0.32	0.24
Item 39	(2)#	1572	0.26	0.07	542 (0.30)	480 (0.23)	0.10	0.02
Item 40	(1)	1651	0.28	0.26	776 (0.43)	358 (0.17)	0.29	0.21

4.10 CTT results for Somali language group

In Somali language group 498 cases were analyzed. There were no missing data. Data below are based on 498 complete cases for 40 data items. A total of 19 items namely Item 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 15, 21, 24, 27, 28, 33, 34, 36 and 39 were found to

be problem items (Table 13). The discrimination indexes of Items 2, 3, 4, 5, 7, 9, 10, 15, 21, 24, 27, 28, 33, 34, and 39 were found below 0.2. There are items with discrimination index close to 0 but not negatively discriminating. Items 3, 6, 9, 10, 15, 27, 28, 33, 34, 36 and 39, were found very difficult to the group and were answered only by less than 30% of the test-takers in each case.

Table 13.Item analysis results for Somali language group

Number Item	Key	Item Correct	Item Diff.	Disc. Index	# Correct in High Grp	# Correct in Low Grp	Point Biser.	Adj. Pt Bis
Item 01	(2)	387	0.78	0.23	139 (0.87)	90 (0.63)	0.25	0.16
Item 02	(4)#	255	0.51	0.09	84 (0.53)	62 (0.44)	0.07	-0.05
Item 03	(3)#	124	0.25	0.17	56 (0.35)	26 (0.18)	0.20	0.10
Item 04	(1)#	200	0.40	0.14	79 (0.49)	50 (0.35)	0.16	0.04
Item 05	(2)#	148	0.30	0.18	63 (0.39)	31 (0.22)	0.16	0.05
Item 06	(3)#	195	0.39	0.22	77 (0.48)	37 (0.26)	0.19	0.08
Item 07	(1)#	97	0.19	0.17	51 (0.32)	21 (0.15)	0.22	0.13
Item 08	(4)	146	0.29	0.28	70 (0.44)	22 (0.15)	0.28	0.17
Item 09	(2)#	139	0.28	0.02	48 (0.30)	40 (0.28)	0.05	-0.06
Item 10	(1)#	135	0.27	0.11	55 (0.34)	33 (0.23)	0.15	0.04
Item 11	(4)	216	0.43	0.43	103 (0.64)	30 (0.21)	0.34	0.23
Item 12	(1)#	253	0.51	0.25	94 (0.59)	48 (0.34)	0.17	0.05
Item 13	(1)#	255	0.51	0.27	102 (0.64)	52 (0.37)	0.21	0.09
Item 14	(4)	153	0.31	0.26	73 (0.46)	28 (0.20)	0.26	0.15
Item 15	(3)#	81	0.16	0.16	43 (0.27)	16 (0.11)	0.20	0.12
Item 16	(2)	370	0.74	0.41	147 (0.92)	72 (0.51)	0.38	0.28
Item 17	(1)	344	0.69	0.34	137 (0.86)	74 (0.52)	0.29	0.19
Item 18	(2)	264	0.53	0.39	117 (0.73)	49 (0.35)	0.32	0.21
Item 19	(4)	130	0.26	0.24	65 (0.41)	24 (0.17)	0.24	0.14
Item 20	(3)	129	0.26	0.29	68 (0.43)	19 (0.13)	0.30	0.20
Item 21	(1)#	208	0.42	0.16	75 (0.47)	44 (0.31)	0.16	0.04
Item 22	(2)	170	0.34	0.40	86 (0.54)	20 (0.14)	0.35	0.25
Item 23	(4)	119	0.24	0.31	65 (0.41)	13 (0.09)	0.32	0.23
Item 24	(4)#	227	0.46	0.25	93 (0.58)	47 (0.33)	0.20	0.09
Item 25	(2)	361	0.72	0.37	141 (0.88)	72 (0.51)	0.36	0.26
Item 26	(2)	185	0.37	0.23	77 (0.48)	35 (0.25)	0.22	0.11
Item 27	(4)#	62	0.12	0.12	32 (0.20)	11 (0.08)	0.14	0.06
Item 28	(3)#	112	0.22	0.09	44 (0.28)	26 (0.18)	0.11	0.01
Item 29	(2)	341	0.68	0.43	136 (0.85)	59 (0.42)	0.38	0.28
Item 30	(3)	331	0.66	0.30	124 (0.78)	67 (0.47)	0.26	0.15
Item 31	(2)	391	0.79	0.21	142 (0.89)	96 (0.68)	0.25	0.16
Item 32	(4)	286	0.57	0.26	109 (0.68)	60 (0.42)	0.23	0.11
Item 33	(4)#	53	0.11	0.10	30 (0.19)	13 (0.09)	0.21	0.14
Item 34	(1)#	75	0.15	0.01	27 (0.17)	22 (0.15)	0.03	-0.06
Item 35	(2)	247	0.50	0.46	113 (0.71)	35 (0.25)	0.37	0.25
Item 36	(3)#	127	0.26	0.17	57 (0.36)	26 (0.18)	0.21	0.11
Item 37	(4)	227	0.46	0.30	96 (0.60)	42 (0.30)	0.27	0.16
Item 38	(3)	126	0.25	0.29	65 (0.41)	16 (0.11)	0.29	0.19
Item 39	(2)#	94	0.19	0.16	48 (0.30)	20 (0.14)	0.22	0.13
Item 40	(1)	100	0.20	0.24	48 (0.30)	9 (0.06)	0.25	0.16

4.11 CTT results for Tigrigna language group

In Tigrigna language group 998 cases were analyzed. There were no missing data. Data below are based on 998 complete cases for 40 data items. A total of 8 items namely Item 5, 9, 12, 13, 18, 26, 28, and 39 were found to be problem items (). Except Item 1, the discrimination indexes of these items were found below 0.2 and unlike to the previous groups there are no negatively discriminating items. Items 5, 12, 13, 28, and 39, were found very difficult to the group and were answered only by less than 30% of the test-takers in each case. Among the problem items Item 26 was the easiest and answered by 90% of the test-takers.

Table 14.Item Analysis Results for Tigrigna Language Group

Number Item	Key	Item Correct	Disc. Diff.	Index	# Correct in High Grp	# Correct in Low Grp	Point Biser.	Adj. Pt Bis
Item 01	(2)	561	0.56	0.28	222 (0.69)	137 (0.41)	0.25	0.16
Item 02	(4)	527	0.53	0.47	251 (0.78)	103 (0.31)	0.41	0.33
Item 03	(3)	388	0.39	0.49	214 (0.67)	61 (0.18)	0.44	0.36
Item 04	(1)	340	0.34	0.37	187 (0.58)	70 (0.21)	0.36	0.28
Item 05	(2)#	201	0.20	0.15	87 (0.27)	42 (0.13)	0.17	0.10
Item 06	(3)	329	0.33	0.21	145 (0.45)	82 (0.25)	0.20	0.11
Item 07	(1)	560	0.56	0.43	258 (0.81)	124 (0.37)	0.38	0.30
Item 08	(4)	799	0.80	0.34	301 (0.94)	201 (0.61)	0.38	0.32
Item 09	(2)#	237	0.24	0.20	112 (0.35)	51 (0.15)	0.19	0.12
Item 10	(1)	546	0.55	0.56	271 (0.85)	96 (0.29)	0.47	0.40
Item 11	(4)	896	0.90	0.23	314 (0.98)	251 (0.76)	0.33	0.28
Item 12	(1)#	74	0.07	0.01	26 (0.08)	22 (0.07)	0.01	-0.03
Item 13	(1)#	209	0.21	0.14	91 (0.28)	47 (0.14)	0.17	0.09
Item 14	(4)	545	0.55	0.47	256 (0.80)	108 (0.33)	0.40	0.33
Item 15	(3)	226	0.23	0.22	112 (0.35)	44 (0.13)	0.22	0.14
Item 16	(2)	570	0.57	0.40	245 (0.77)	123 (0.37)	0.36	0.28
Item 17	(1)	586	0.59	0.50	271 (0.85)	114 (0.34)	0.44	0.37
Item 18	(2)#	320	0.32	0.14	129 (0.40)	87 (0.26)	0.13	0.04
Item 19	(4)	382	0.38	0.35	185 (0.58)	77 (0.23)	0.34	0.26
Item 20	(3)	496	0.50	0.51	248 (0.78)	88 (0.27)	0.44	0.37
Item 21	(1)	330	0.33	0.30	166 (0.52)	72 (0.22)	0.30	0.22
Item 22	(2)	472	0.47	0.48	230 (0.72)	80 (0.24)	0.42	0.34
Item 23	(4)	428	0.43	0.21	167 (0.52)	103 (0.31)	0.21	0.13
Item 24	(4)	356	0.36	0.27	151 (0.47)	68 (0.20)	0.22	0.14
Item 25	(2)	829	0.83	0.40	317 (0.99)	195 (0.59)	0.48	0.43
Item 26	(2)#	900	0.90	0.24	317 (0.99)	249 (0.75)	0.40	0.36
Item 27	(4)	321	0.32	0.24	150 (0.47)	77 (0.23)	0.23	0.15
Item 28	(3)#	237	0.24	0.03	83 (0.26)	76 (0.23)	0.05	-0.02
Item 29	(2)	259	0.26	0.21	121 (0.38)	55 (0.17)	0.21	0.13
Item 30	(3)	803	0.80	0.28	293 (0.92)	210 (0.63)	0.37	0.31
Item 31	(2)	826	0.83	0.31	302 (0.94)	212 (0.64)	0.40	0.34
Item 32	(4)	623	0.62	0.25	235 (0.73)	160 (0.48)	0.25	0.16
Item 33	(4)	318	0.32	0.24	148 (0.46)	73 (0.22)	0.24	0.16
Item 34	(1)	573	0.57	0.37	245 (0.77)	133 (0.40)	0.33	0.24
Item 35	(2)	599	0.60	0.47	271 (0.85)	126 (0.38)	0.40	0.32
Item 36	(3)	448	0.45	0.35	213 (0.67)	104 (0.31)	0.35	0.27
Item 37	(4)	661	0.66	0.30	257 (0.80)	167 (0.50)	0.27	0.19
Item 38	(3)	478	0.48	0.46	228 (0.71)	84 (0.25)	0.43	0.35
Item 39	(2)#	151	0.15	0.23	97 (0.30)	24 (0.07)	0.33	0.27
Item 40	(1)	781	0.78	0.32	294 (0.92)	200 (0.60)	0.35	0.29

4.12 Summary test statistics for the whole data and by language groups

The summary statistics in Table 15 presents the average difficulty level, discrimination index, the reliability of the test and others at the test level by language groups. Looking at the difficulty level it ranges from 0.35 in English to 0.48 in Tigrigna indicating that the English language group test-takers performed less than the others. In terms of the average discrimination index the test in Tigrigna was highly discriminating and the test in Somali was discriminating the least. This indicates that the test in Somali was difficult to both upper and lower group examinees.

Table 15. Summary test statistics based on the whole data and by language group

Group	All	Afan Oromo	English	Somali	Tigrigna
Mean Item Difficulty	**0.386**	**0.442**	**0.350**	**0.395**	**0.481**
Mean Item Discrimination	**0.300**	**0.249**	**0.274**	**0.238**	**0.310**
Mean Point Biserial	0.283	0.245	0.276	0.232	0.309
Mean Adj. Point Biserial	0.202	0.152	0.190	0.129	0.235
KR20 (Alpha)	0.713	0.628	0.691	0.552	0.758
KR21	0.691	0.538	0.672	0.465	0.698
SEM (from KR20)	2.886	2.779	2.856	2.799	2.747
Potential Problem Items	10	12	10	19	8
High Grp Min Score	18	20	16.000	18.000	22.000
Low Grp Max Score	12	15	11.000	13.000	16.000
Number of Examinees	9533	2038	5999	498	998
Standard Deviation	5.392	4.557	5.136	4.182	5.587
Mean Score	**15.44 (38.6%)**	**17.68 (44.2%)**	**14.02 (35.0%)**	**15.79 (39.5%)**	**19.22 (48.1%)**

In terms of reliability the alpha value for Tigrigna 0.758 was the highest and in the acceptable range while that of Somali is the least and not acceptable. The reliability for Afan Oromo and English language groups were also low.

Chapter. 5 Summary, conclusions and recommendations

5.1 Summary

The National Learning Assessment (NLA) in Ethiopia is a large scale national survey conducted by the National Examination Agency under the Ministry of Education. It was first conducted in 2000 and has been repeated every three or four years since then. In Grade 8 five subjects namely biology, chemistry, physics, English and mathematics were tested. This study investigated Item and Test Analysis of Grade 8 Biology Test of the Ethiopian Third National Learning Assessment (ETNLA). A total of 10,795 students sat for the biology test in 2007, of these 9,552 were used for the study. The test was originally prepared in English and was then translated into three language versions (Afan Oromo, Somali and Tigrigna). The main purpose of the study was to see how the items worked across language groups.

All the necessary measures were taken to make the test valid and reliable during its development and translation. During the last decades, test adaptations and translations have become prevalent because of an increase in national assessment testing in multiple languages, and a growing concern to test students in their first language and Ethiopia is not an exception. The comparability of test results across different language versions of these tests is at the core of the validity of interpretations in these assessments.

As part of the test development process, analysis of the items is a crucial part. Two prevailing methods, both with strengths and weaknesses, are predominantly used. In the Classical Test Theory, its ease of use and adaptability in analyzing practically all kinds of tests renders it a popular choice. However, its strong dependence on the kind

of sampling required often limits its applicability. Hence, CTT developed tests would see the need for bigger sampling every now and then which in the long run renders it expensive. On the other hand, the emerging Item Response Theory (IRT) seems to have found a way to avoid the pitfalls of CTT. It is said to be sample free or sample independent. The only drawback is the cumbersome statistical analysis required which other test developers would shy away from. Nevertheless, IRT is slowly gaining momentum in the field of psychology (Andrade, Tavares and Valle, 2000).

A two Parameter Logistic Model (2PLM) based on Item Response Theory was used to investigate latent traits and the main statistics generated were IRT ability score, difficulty level and discrimination index. Item Characteristic Curves (ICCs) were plotted for all 40 items by language groups. Language groups were compared using one-way anova and recursive partitioning analysis. Classical Test Theory procedure was also used to generate item and test statistics and results were compared with that of IRT findings.

The ICC varied a lot across language groups and the number of problem items by language group based on CTT was: Somali (19), Afan Oromo, (12), English (10) and Tigrigna (8).The highest test score 22 was for Tigrigna, followed by 20 for Afan Oromo. The English language group students scored the least (15). The performance of Somali language group students were almost equal to that of English group ones. Nevertheless, considering the number of problem items it is possible to say the former performed better. The finding show that there were items which did not work the same way across the four language groups which make them as language differential item functioning suspects.

5.2 Conclusions

Item analysis is usually used to answer questions such as: How difficult is the item? How well does the item discriminate between high and low achievers? And how effective is each distracter? The result of item analysis can help test developers to prepare valid and reliable tests. Furthermore, such kind of information is important to prepare guidelines about test development and conducting assessments.

Response to the test items should be analyzed for several reasons, the most important of which is eventually improvement of the items and consequently the test. The difficulty level and the discriminating power of each item are the keys to item improvement. Item analysis can also lead to improve instruction through identification of weaknesses in the examinees as a group, in instructional methods, or in the curriculum. When it is done across subgroups, item analysis addresses issues of fairness.

The findings of this study show that test and item statistics showed substantial variations across the four language groups. Items which were found difficult to one language group were found easy to the others and vice versa. Students tested in their mother tongue except those tested in Somali performed better than those tested in English.

Based on the IRT ability scores 20% of the variations in the achievement scores was due to language differences. The variation explained just using a single variable is very high and it shows how important is the issue of language. With the advent of powerful statistical packages shying away from using IRT procedures is no more necessary though CTT is easy to use and informative.

Based on CTT results the number of problem items was highest for Somali language group test-takers and least for those of Tigrigna ones.

The study showed that relying heavily on experts' judgments and following appropriate procedures during translation by its self cannot ensure the validity of individual items in particular and the test as a whole. The IRT person-item dual plot showed that the items for different ability groups across language groups showed wide variations. In order to make fair and unbiased judgments based on the scores a more robust technique should be employed.

5.3 Recommendations

The findings showed that there is a substantial variation in the psychometrics statistics of the items hence it is recommended that in future undertakings:

- It is very import to check the consistency of the items across different groups
- Piloting should be followed with detail item and test analysis
- Test scores based on CTT alone cannot give the clear picture and the sample and item independent IRT procedures should be employed
- The agency should produce separate reports for each subject and these reports should contain detail test and item statistics across subgroups including language groups

References

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology*, 77, 598-614.

Adedoyin, O. O., Nenty, H.J, and Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review, 3* (2), 83-93.

Algina, J. and Crocker, L. (1986).*Introduction to Classical and Modern Test Theory*, Holt, Rinehart and Winston.

Allalouf, A., Hambleton, R., and Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36,*185-198.

Anastasi, A. and Urbina, S. (2002).*Psychological testing*. Prentice Hall: New York.

Andrade, D. F., Tavares. H. R., and Valle, R. C. (2000). *Teoria da Respostaao Item: conceitos e aplicações*. São Paulo: ABE.

Angoff, W. (1993).Perspective on differential item functioning methodology. In P.W. Holland and H.Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Baker, F. B. (2001). *The Basics of Item Response Theory.* USA: Heinemann.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical Theories of*

Mental Test Scores (pp. 397 - 472). Reading, MA: Addison-Wesley Publishing.

Bolt, D., Froelich, A., Habing, B., Hartz, S., Roussos, L., and Stout, W. (1999). An *applied and foundational research project addressing DIF, impact, and equity with applications for ETS test development* ETS Research Report (RR-03-06). Princeton, NJ: Educational Testing Service.

Camilli, G. and Shepard, L. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Clark, L.A. and Watson, D. (1995).Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.

Clauser, B.E., and Mazor, K. M. (1998). Using statistical procedures to identify Differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Crocker, L., andAlgina, J. (1986).*Introduction to Classical and Modern Test Theory.* Fort Worth: Harcourt Brace Jovanovich College Publishers.

Dorans, N. J., and Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel And Standardization. In P. W. Holland and H. Wainer (Eds.) *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Educational testing service fairness review guidelines (2003). Princeton, NJ: Educational Testing Service.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-385.

Florida Department of Education (2005). *FCAT Handbook, A Resource for Educators*. Florida Department of Education, Tallahassee, Florida.

GEQAEA (2007).*Ethiopian Third National Learning Assessment of Grade 8 Students*. Addis Ababa.

Gierl, M. J., Khaliq, S. N. and Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. Paper presented at the Symposium entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Quebec, Canada.

Glas, C.A.W. *et al.* (2003). *Educational Evaluation, Assessment, and Monitoring, A Systemic Approach,* Sweets and Zeitlinger Publishers

Hambelton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38*, 60-65.

Hambleton, K.R, Swaminathan, H. and Rogers, J.H.(1991). *Fundamentals of Item Response Theory*, Sage Publications, Inc.

Hambleton, K.R. and Swaminathan, H.(1985).*Item Response Theory, Principles and Applications,* Kluwer Nijhoff Publishing.

Hambleton, R. K., and Swaminathan, H. (1995).*Item Response Theory: Principles and Applications*. Norwell, MA: Kluwer Academic Publishers.

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

Hauser, C., and Kingsbury, G. (2004).*Differential item functioning and differential test functioning in the "Idaho Standards Achievement Tests" for spring 2003.* Lake Oswego, OR: Norwest Evaluation Association.

Holland, P. W., and Thayer, D. T. (1988).Differential item functioning and the Mantel- Haenszel procedure. In H. Wainer and H. I. Braun(Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Impala, J. C., and Place, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement,35*, 69-81.

Kaplan, R.M. and Accuse, D.P.(1997).*Psychological testing: Principles, applications and issues*. Pacific Grove: Brooks Cole Pub. Company.

Lagerfeld, T. E. (1997). Test fairness: Internal and external investigations. *Educational Measurement: Issues and Practice, 16,* 20-26.

Lord, M. F. and Novick, R. M. (1968). Statistical Theories of Mental Test Scores, Addison- Wesley Publishing Company.

Lord, M. F. and Novick, R.M.(1968). *Statistical Theories of Mental Test Scores*, Addison- Wesley Publishing Company.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-331.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology, 3*, 1 –18.

Nuked, C. (2002) A Study of Raven Standard Progressive Matrices test's item measures under classic and item response models: An empirical comparison. *Ankara University, Journal of Faculty of Educational Science, 35* (1-2), 71-79.

Penfield, R. D. and Camilli, G. (2007). "Differential item functioning and item bias". In C.R. Rae and S. Sinhala (Vol. Eds.), *Handbook of statistics: Vol. 26* (pp. 125 – 167). Elsevier.

Settler, S., and Qualls, A. L. (2000).Examination of differential item functioning on a standardized achievement battery with limited English proficient students. *Educational and Psychological Measurement, 60*(4), 564–577.

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

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485-518.

Thissen, D., Steinberg, L., and Wainer, H. (1998). Use of item response theory in the study of group threshold differences in trace lines. In Wainer, H. and Braun, H., editors. *Test validity*. Hillsdale, NJ: Lawrence Erlbaum, 147-69.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. University of North Carolina at Chapel Hill.

Van der Linden, A., and Humbleton, R. (1980). *Introduction to scaling*. New York: Wiley.

Wainer, H., Sireci, S. G., and Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197-219.

Yen, M.W and Allen, J. M.(1979). *Introduction to Measurement Theory*, Brooks/Cole Publishing Company.

Zenisky, A. L., Hambleton, R.K., and Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1-2), 61–78.

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

Appendix Test blue print

GRADE	UNIT	CONTENT	PERIOD	BEHAVIOR				CONTENT
				K	C	APP	ASE	TOTAL
7	1	Single Celled Organisms • Biology As A Subject • Single Celled Plants • Single Celled Animals • Bacteria	20	2	2			4
7	2	The Cell • Cell Theory • Structures o Plant Cells o Animal Cells • Cells And Tissues • Organs And Organ Systems	10	1	1			2
7	3	Habitats • Different Types Of Habitats • Studying Habitat • Community And Succession • Food Relationships	12	1		1		2
	4	Non-Flowering Plants • General Features • Green Algae • Fungi • Lichens • Mosses • Ferns	12	1	1			2
7	5	Insects • Characteristics • External Body Structures • Metamorphosis • Life Histories o House Fly o Mosquito o Locust • Social Insects o Honey Bee	13	1	1			2
7	6	Human Biology and Health • Muscles and The Skeletal System • Food and Nutrition • Digestion • Circulatory System • Excretion	33	3	3	2		8

GRADE	UNIT	CONTENT	PERIOD	BEHAVIOR				CONTENT
				K	C	APP	ASE	TOTAL
8	1	Human Biology and Health • Sense Organs o Eye o Ear o Tongue o Nose o Skin • The Nervous System • The Endocrine System • The Reproductive System	32	3	3	2		8
	2	Humans and Diseases • Plant Diseases • Human Diseases o Causes o Transmission and Prevention • Defense Against Diseases • Community Hygiene • Curing Diseases	21	2	2			4
8	3	Flowering Plants • The Organs of Flowering Plants • Monocots and Dicots • Reproduction of Flowering Plants o Asexual o Sexual • Seeds and Fruits • Storage Organs	17	2	1			3
8	4	Photosynthesis • Mechanism of Photosynthesis • Testing for Starch • Comparison of Photosynthesis with Respiration	12	3	2			5
Total			**207**	**19**	**16**	**5**		**40**

DECLARATION

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Name: _____

Signature: _____

Date of Submission: _____

This thesis has been submitted for examination with my approval as a university advisor.

Name: _____

Signature: _____

Date: _____