Master Thesis

IRT in Item Banking, Study of DIF Items and Test Construction

(Item Response Theory in Item Banking, Study of Differentially Functioning Items and Test Construction)

Gembo Tshering
s0129836
M.Sc Student

Master of Science-Track Educational Evaluation and Assessment
Educational Science and Technology
University of Twente
The Netherlands

Bhutan Board of Examinations
Ministry of Education
Royal Government of Bhutan
Bhutan

Netherlands Fellowship Programmes
Netherlands Organization for International Cooperation in Higher Education (NUFFIC)
Ministry of Education, Culture and Science
The Netherlands

University of Twente
Educational Science and Technology
The Netherlands

CITO
National Institute of Educational Measurement
The Netherlands

This page is dedicated to the above organizations for giving me an opportunity to develop academically and acquire the capacity to interact with knowledge and wisdom available within the vast envelope of education.

Ms Angela Verschoor is an expert in CAT, but I soon learned that I would not have a share of knowledge from her only because UIBTERV is a paper and pencil tests project. Dr. Bas Hemker is a psychometrician at CITO and UIBTERV is one of his projects. Dr. Bas Hemker not only spared UIBTERV for me, but he also consented to be my external supervisor at CITO who is due to be involved almost daily with me on UIBTERV. It is with constant maneuvering and guidance from Dr. Bas Hemker that I discovered what I have been looking for in UIBTERV and made my work a success.

Dr. Bas Hemker led me to a couple of meetings on entrance test for English Reading Comprehension and Vocabulary test with test experts Ms. Marion Feddema and Ms. Noud van Zuijlen, where I learnt how the test statistics are used for making decision about the prospective examinees.

OPLM, item calibration, study of differential item functioning, test construction and extending item bank have been continuous lessons imparted to me by Dr. Bas as I ploughed through UIBTERV. Indeed, it is his continuous support, interest and confidence in my progress with UIBTERV that I realized I could complete the project as per the deadline set in the contract.

I also continuously enjoyed the goodwill of CITO through other personalities. “A theoretical knowledge is never as beautiful as a practical knowledge” is a comment from Prof. Dr. Piet Sanders, Head of the Department of Psychometrics, CITO, which echoed continuously and triggered a daily quest that I would claim as a realization at the end of a hard day work. Dr. Timo Bechger’s demonstration of Newton Raphson Estimation procedure as used for MML estimation with MathCad is a pleasure hard to forget.

My friends Mr. Wouter Toonen, and Dr. Huseyin Yildrem have been wonderful friends. Although, each one of us was always conscience stricken by fast evading time, we would occasionally manage to turn around and laugh over small talks.

Last but not least, I would have been extraordinarily forgetful if I did not mention my association with the second mentor DR. Hans Vos. Dr. Hans Vos has been a constant source of inspiration to me. His interest in me led me to an additional experimental success at CITO through conducting an experiment on *Standard Setting by using Bookmark method* in collaboration with Dr. Maarten de Groot, psychometrician. Dr. Hans Vos has read the complete thesis and I once again say thank you to him for his valuable feedback and comments.

Finally I would like to thank everyone who has been a part of this thesis. My thanks are due to (a) NUFFIC for sponsoring me the master of science (M.Sc) degree course vide fellowship award letter NFP-MA.05/1518\GW.07.05.250/fl\file number 022/05\dated June 9, 2006, (b) University of Twente, Educational Science and Technology for being home and family abroad while pursuing M.Sc. and (d) National Institute of Educational Measurements, CITO, for contracting the project and also sponsoring my daily involvement in it.

Gembo Tshering
Educational Science and Technology
University of Twente
Enschede
The Netherlands

13 July 2006

Contents

Chapter 1

Using Item Bank in making Tests for English Reading and Vocabulary Constructs

1.0.0 Introduction

The project *Using Item Bank in making Tests for English Reading and Vocabulary* (UIBTERV) consists of a series of continuous stage by stage tasks. The stages that UIBTERV went through were (a) writing test items for English reading and vocabulary constructs for Lower Secondary Education of the Netherlands, (b) assembling the test items into thirteen test booklets with each booklet containing seventy tests items, (c) pre-testing the test booklets with sample schools, (d) correction and scoring of the pretested booklets and (e) construction of data banks for the scores obtained from the pretests.

My role in UIBTERV begins from the last stage, i.e., stage e. Although my role begins somewhere from the middle and will end somewhere short of touching the end of UIBTERV, it is important to understand UIBTERV from its holistic perspective to have an idea of an immensity of stake it carries against thousands of Dutch students.

In this chapter, a brief description of the Dutch education system will be presented and UIBTERV will be described in terms of its goals, current situation and future line of plans in the light of Dutch Secondary Education. Finally, this chapter will complete with short tour of the subsequent chapters.

1.1.0 The Dutch Education System

The education system of the Netherlands is composed of different levels. The different levels are Primary Education, Special Primary Education, Lower Secondary Education, Pre- University Education, Senior General Secondary Education, Pre-Vocational Secondary Education, Vocational Training Programme, Special Secondary Schools, University Education, Higher Professional Education, and Senior Secondary Vocational Education. The relationships among the different levels and the direction of movements form one level to another level are displayed in figure 1.1.0.1 (CITO, 2005, p.7).

The Primary Education in the Netherlands spans a duration of eight years. The children are allowed to begin school at the age of four and they are legally obliged to be in the schools at the age of five. Of the eight years, the last seven years are compulsory. When the children graduate from the Primary Education, they are given a recommendation on the type of secondary education that they should pursue. The recommendation is based on the results of the four different kinds of tests conducted in four areas, viz., (a) language, (b) arithmetic/mathematics, (c) study skills and (d) world orientation. The teachers also play very important role in enhancing the validity of the recommendation by providing additional information about their students. The Primary Education has also Special Primary Education for the students with learning and behavioral problems.

The Dutch children begin their Secondary Education at the age of twelve and may continue up to the age of eighteen. The Secondary Education has various intra-Secondary Education levels. The first and the second years of the Secondary Education form the Lower Secondary Education. The Lower Secondary Education has different sections of schooling, viz., VWO, HAVO, GT, KB, BB and BB+. Students are classified into these sections based on their performance in the Primary Education, however. The students have freedom to change their sections based on their study calibers. After successfully completing the year 2 of the Lower Secondary Education, the students can move to year 3 and may continue up to year 6 of the Secondary Education. The years 3 through 6 of the Secondary Education consist of different sections of schooling with the sections having more homogeneous groups of students. The students from VWO pursue Pre-University Education, the students from HAVO pursue Senior General Secondary Education and the students from GT, KB, BB, BB+ pursue Pre-Vocational Secondary Education. The Secondary Education also has Special Secondary Education for the students having learning and behavioral problems.

The VWO curriculum prepares students for research oriented university education, the HAVO curriculum prepares them for professional university education and the Pre-Vocational Secondary Education curriculum prepares them for Senior Secondary Vocational Education.

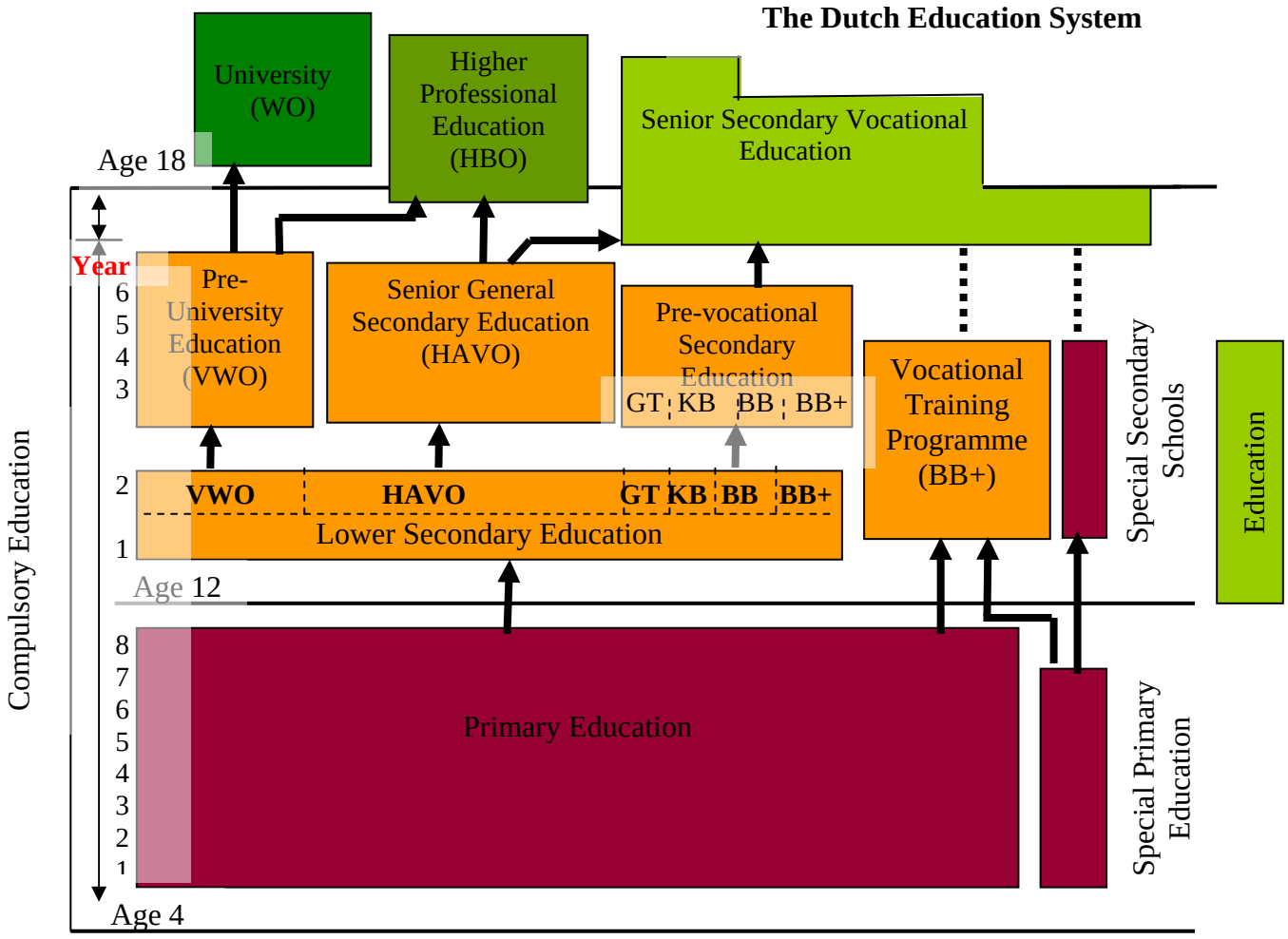

Figure 1.1.0.1: The Dutch Education System

1.1.1 Student Monitoring System

The movements of students from Primary Education to Secondary Education and the types of schooling they will pursue are monitored by the Student Monitoring System. Generally the Student Monitoring System is implemented in Primary Education and in Secondary Education. The professional body responsible for developing, managing, executing and reporting the matters relevant to the Student Monitoring Systems is the National Institute of Educational Measurement, CITO, located at Arnhem in the Netherlands. The National Institute of Educational Measurement does everything concerning educational assessment, measurement and evaluation in the Netherlands and to a certain extent its expertise sail abroad as well. The readers may like to visit www.cito.com for more discoveries and information about the National Institute of Educational Measurement, CITO, Arnhem, the Netherlands.

The Student Monitoring System is highly psychometric based on quantitative research with elements of longitudinal design. At both Primary Education and Secondary Education levels, the Student Monitoring System consists of entrance tests, follow-up tests and advisory tests. The contents of the

student monitoring systems are different for Primary Education and Secondary Education because they have structural differences. The Student Monitoring System of Primary Education is not within the scope of the thesis and also the scope of the thesis does not warrant a detailed description of the Student Monitoring System of the Secondary Education. In case of the Student Monitoring System of the Secondary Education, a detailed description of the role of the Department of Psychometrics at CITO in producing entrance tests, follow-up tests and advisory tests will be presented throughout the thesis.

The Student Monitoring System of the Secondary Education seeks to

1. measure an individual student's competence.
2. measure student's competence relative to the fellow students.
3. measure the progress of an individual student.
4. measure the progress of a student with reference to the fellow students.
5. advise student on the choice of the types of schools he/she could pursue in the third year of the Secondary Education.

The five goals of the Student Monitoring System of the Secondary Education link students with teachers by presenting answers to the following questions:

- How much progress is made by my students?
- Is the progress made by my students sufficient?
- Is there a standstill or deterioration in the students' development and what can I do about it as a teacher?
- Is the subject matter offered by me as a teacher adequately geared to the level of the students?
- Which student needs extra help and attention?
- Do I have to address or change my didactic approach as a teacher?
- Are there any parts of the educational programme in need of improvement?
- What are the positions of my students with respect to the students of other classes and the students of other schools?

To achieve the five goals and to get answers to the questions, the Student Monitoring System of the Secondary Education uses three tests, viz., (a) entrance tests, (b) follow-up tests and (c) advisory tests.

In a nutshell, the School Monitoring System purports to (a) help teachers monitor their students' development, (b) provide tools to help students decide on the type of schooling they should choose after successfully completing the Lower Secondary Education and (c) monitor the quality of the educational process. UIBTERV has direct relevance to the Student Monitoring System of the Secondary Education.

1.2.0 UIBTERV

UIBTERV pervades all stages of the Student Monitoring System of the Secondary Education of the Netherlands. UIBTERV is implemented in the Lower Secondary Education. The Student Monitoring System for Secondary Education has four areas of test, viz., (a) Dutch Reading Comprehension, (b) English Reading Comprehension,(c) Mathematics and (d) Study skills. The test in each of these four areas has three phases known as (a) entrance test at the start of first year, (b) test after the first year of the secondary school (alias: Follow up Test) and (c) test after the second year of the Secondary School (alias: Advisory Test). UIBTERV is related only to English Reading Comprehension with addition of English Vocabulary domain. Table 1.2.0.1 shows the summary of the different tests involved in the Student Monitoring System of the Secondary Education.

Test Area	Three Tests		
	Entrance Test	Test after 1st Year	Test after 2nd Year
Dutch Reading Comprehension	Prepared	Prepared	Prepared
English Reading Comprehension	UIBTERV	Prepared	Prepared
Mathematics	Prepared	Prepared	Prepared
Study Skills	Prepared	Prepared	Prepared

Table1.2.0.1: Shows test areas and types of tests offered by *The Student Monitoring System*.

In table 1.2.0.1, prepared indicates that the tests were ready for administration. Where UIBTERV is concerned, the tests are to be prepared. Each test in table 1.2.0.1 has three versions corresponding to three difficulty levels. For instance, an entrance test in English Reading Comprehension has three different tests of three different difficulty levels, viz., (a) an Easy and Average Easy Test, (b) an Average Easy and Average Difficult Test and (c) an Average Difficult and Difficult Test. Table 1.2.0.2 shows this relationship.

The combination of (a) easy items and average easy items makes up an Easy and Average Easy Test for the students studying in BB+/BB section of the Secondary Education, (b) average easy items and average difficult items makes up an Average Easy and Average Difficult Test for the students studying in KB/GT section of the Secondary Education and (c) average difficult items and difficult items makes up an Average Difficult and Difficult Test for the students studying in HAVO/VWO section of the Secondary Education.

The three versions of the test are overlapped by using common items. The test for BB+/BB and the test for KB/GL are overlapped by using average easy items as the common items. The test for KB/GT and the test for HAVO/VWO are overlapped by using average difficult items as the common items. The common items link the three versions of the test which is a necessary condition to place them on a common scale for comparative studies of their results.

Target Population	Entrance Test			
	Easy	Average Easy	Average Difficult	Difficult
BB+/BB				
KB/GT				
HAVO/VWO				

Table 1.2.0.2: Shows how Entrance Test is divided into three tests for three levels of secondary education.

1.2.1 UIBTERV Data

UIBTERV DATA is collected from the English Reading Comprehension and Vocabulary pretests. The pretests are carried out to gather information about the test items and examinees. The information about the items comprise of how the items functioned in the tests, how the items functioned across different groups of populations and how the items contributed to the overall goals of the tests. The information about the examinees comprise of how they responded to the items in a test and how their demographic background affected their performance in the test. The information can be generated by analyzing the data from the pretests. The information obtained from the analyses can be used in selecting good items for the real test and designing test suitable for different groups of examinees.

UIBTERV DATA consists of thirteen dichotomized datasets- each dataset coming from its own booklet. Each booklet consists of two tasks, each with 35 items. Thus each booklet contains 70 items.

The test booklets were administered to 1280 students. The data are joined in a complete dataset. A portion of data from the data file *EngEntNw.DAT* is shown in table 1.2.1.1.

```
19828 2218 01 J 19085 0 0 39    1   1 1   1 10011000101110011011100100101011111010111011011001010101001011001110 10
19828 2220 01 J 08045 1 0 51    1   1 1   1 11101100100111111111010110111111011111110111011101001111001001111111110 10
19828 2190 01 J 06125 1 0 59    1   1 1   1 11111100110110111111110111111111111111100111011101110111101111111101111111 10
19828 2189 01 J 09046 1 0 62    1   1 1   1 11111101110111111111111111110111101111110111111110111111101111110111111 10
. . .
. . .
. . .
. . .

30442 0163 02 J 07024 6 0 45    2   2 2   2 00001101110001110010111101101000010111011111111111011110101011101011111 10
22906 0096 02 J 25066 1 0 53    2   2 2   2 11011001110101101011111011111111010111011011101111111010110111101111111 10
30442 0166 02 J 21025 1 0 56    2   2 2   2 11101011101101011111111111111101110111111111111011110011011011110111111 10
22906 0093 02 J 29115 1 0 51    2   2 2   2 11111010001001110101111111111011100111110111101011011101100111011111 10
. . .
. . .
. . .

30442 0161 02 J 18074 5 0 62    2   2 2   2 11111111011111110111111111101101100111111111111111101101111111111111111 10
22906 0095 02 J 05016 1 0 59    2   2 2   2 11111111101101111111111111111101110111111111101111011010101111001111111 10
30442 0162 02 M 03035 3 0 36    2   2 2 15 01100001001001101100110101111000110111101011100010001010110011110010 10
```

Table 1.2.1.1: Excerpts from UIBTERV data file *EngEntNw.DAT*.

The data has the following variables with the positions specified against them:

Position	Variables
1-5	Case number
1-11	Student number
14-15	Old booklet number
18	Gender (J=jongen=boy, M=meisje=girl)
21-25	Date of birth (ddmmY, with Y=Y+1987 (so 1=1988,2=1989, etc)
28	Language at home (1=Dutch, 2= Turkish, 3= Arabic, 4=Surinam, 5=English, 6=Any other)
31	School level (1=BB+, 2=BB, 3=KB, 4=GL, 5=HAVO, 6=VWO)
42-43	Correct key (new booklets)
46-47	Mm= many missing=last 35 items missing, m=missings>10 missing
51-52	New booklet
	1-13=complete booklet 1-13 (according to correct key
	14-26= first half (35 items) booklet 1-13 (according to correct key)
	27-28= second half of booklet 1 and booklet 3 (according to correct key)
54-55	DIF booklets School level
58-59	DIF booklets Gender
61-130	Response on items

Table 1.2.1.2: Variables and their positions.

1.2.2 Problems in UIBTERV

As described before, UIBTERV is an important and integral part of the Student Monitoring System of the Secondary Education. The tests in three areas, viz., (a) Dutch Reading Comprehension, (b) Mathematics and (c) Study Skills had been processed and are ready for administration and large item banks were also constructed for future use in assembling similar tests.

A Follow-Up Test and an Advisory Test in English Reading Comprehension are ready for administration. An Entrance Test in English Reading Comprehension and Vocabulary is due to be prepared from the pretested test booklets and the unused items have to be banked for future use in assembling similar tests. The situation elevates a platform to delineate problems in UIBTERV.

The first challenge is to calibrate the pretested items from the thirteen test booklets, each test booklet having 70 test items with some anchor items. The anchor items relate the items from different booklets to each other and make them comparable. The second challenge is to assemble entrance tests for Secondary Education for three populations, as shown in figure 1.2.0.2. The third challenge is to develop norm tables for the entrance tests. The fourth challenge is to successfully link the new items to the old item bank. The old item bank has the items from the Follow-Up Tests and the Advisory Tests on English Reading Comprehension and Vocabulary.

By linking items from Entrance Tests with other items available in the old item bank, the items will be placed on the common scale which will make the comparative study of the results from three tests on English Reading Comprehension and Vocabulary possible. The comparative study across all three tests will make the monitoring of student's learning progress possible.

1.2.3 Goals of UIBTERV

The goals of UIBTERV are to

(1) prepare test specifications.
(2) prepare item pool.
(3) field test the items.
(4) calibrate the field tested items.
(5) develop global norm.
(6) assemble test items.
(7) develop local norms.
(8) extend the old item bank by adding items from the field tested items.
(9) publish entrance test for lower secondary education.

The goals 1, 2 and 3 were successfully completed before I took up UIBTERV and goal 9 is beyond the scope of my contract with CITO, Arnhem, the Netherlands. Therefore, I will elaborate on goals 4, 5, 6, 7 and 8 in chapters 8 through 12.

1.3.0 Chapters in the Thesis

The thesis has 12 chapters. Chapter 1 offers an introduction to the rest of the chapters and their contents. Chapters 2 and 3 offer a quick review of the Classical Test Theory (CTT). As far as possible, the theoretical aspects of CTT are transformed into practical source of information by illustrating them with extracts from the analyses performed by using CTT in UIBTERV. Chapter 4 describes norm reference table and its use in tests. Chapter 5 focuses on the different assumptions of IRT. Chapter 6 presents the One-Parameter Logistic Model. Chapter 7 describes item and test information functions.

While I tried to illustrate the contents of chapters 4 through 7 with extracts from the analyses performed by using IRT in UIBTERV, I must state that some of them could not be readily illustrated with information from UIBTERV, however. The chapters build a concrete stage for orchestrating the connection between theory and application.

Chapters 8 through 11 comprise of my day to day work report on UIBTERV. The report is built upon the hands-on experiences accrued in the course of analyzing UIBTERV data by using both CTT and IRT. While the same purpose of CTT and IRT at times make them equally competent, they are still distinguishable. Based on their qualifying attributes befitting the needs of a test designer, CTT and

IRT are used on the basis of as and where their functions are optimal, complementary and supplementary. Chapter 12 presents some discussions on UIBTERV.

1.3.1 Summary

UIBTERV has gone through (a) writing test items for English Reading and Vocabulary constructs for Lower secondary Education, (b) assembling the test items into 13 test booklets, (c) pre-testing the test booklets, (c) correction and scoring of the pretested booklets and (e) construction of data banks for the scores obtained from the pretests.

The Student Monitoring System is highly psychometric based on quantitative research with elements of longitudinal design. The Student Monitoring System purports to (a) help teachers monitor their students' development by looking at their performance in the tests, (b) provide tools to help students decide on the type of schooling they should choose after successfully completing the Lower Secondary Education and (c) monitor the quality of the educational process. The Student Monitoring System of Secondary Education has four areas of test, viz., (a) Dutch Reading Comprehension, (b) English Reading Comprehension,(c) Mathematics and (d) Study skills.

UIBTERV's goals are to (a) prepare test specifications,(b) prepare item pool, (c) field test the items, (d) calibrate the field tested items, (e) develop global norm, (f) assemble test items, (g) develop local norms, (h) extend the existing item bank by adding items from the field tested items and (i) publish entrance test for lower secondary education.

When items from different tests on English Reading Comprehension and Vocabulary are linked, the items are placed on common scale. Items from different tests with common scale will make the comparative study of the results from different tests possible, meaning that a student's learning progress can be monitored.

1.3.2 References

The Education System in the Netherlands, retrieved on 6 March 2006 from:
http://www.nuffic.nl/pdf/dc/esnl.pdf.

About Cito National Institute for Educational Measurement: May 2005, Cito, Arnhem, Netherlands.

Chapter 2

A Review of Classical Test Theory

2.0.0 Introduction

Item writing and building test precede item bank construction. To construct item bank, items will have to be pilot tested and analyzed by using item response theory or classical test theory or both. In chapter one it was noted that the goal of the project is to extend the old item bank and construct an Entrance Test to measure English Reading Comprehension and Vocabulary constructs of the Dutch students in Lower Secondary Education. Therefore, a review of classical test theory (CTT) is made in this chapter. Emphasis is made on the concepts of the CTT which are directly used for analyzing the data from the pilot tested tests, building item banks and constructing tests.

2.1.0 Classical Test Theory

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.

2.2.0 Assumptions of Classical True Score Theory

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

2.2.1 Assumption 1

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

$$X=T+E. \qquad (2.2.1.1)$$

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.

2.2.2 Assumption 2

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

$$\xi(X) = T\,. \qquad (2.2.2.1)$$

Equation 2.2.2.1 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 & Crocker (1986, p. 109) define the true score as the mean or expected value of a random variable. For a random variable *X* with finite number of discrete values, the expected value of *X* is defined as

$$\mu = \sum_{k=1}^{k} X_k p_k \,, \qquad (2.2.2.2)$$

where X_k is the k^{th} value the random variable can assume, and p_k is the probability of that value. When an observed test score is considered as a random variable, X_j, the true score for examinee *j* is defined as

$$T_j = \varepsilon X_j = \mu_{X_j} \,. \qquad (2.2.2.3)$$

2.2.3 Assumption 3

Assumption three states that the error scores and the true scores obtained by a population of examinees on one test are uncorrelated. Mathematically, this assumption is expressed as

$$\rho_{ET} = 0 \,, \qquad (2.2.3.1)$$

where ρ_{ET} is the correlation between error scores and true scores.

2.2.4 Assumption 4

Assumption four states that the error scores on two different tests are uncorrelated. Mathematically, this assumption can be expressed as

$$\rho_{E_1E_2} = 0 \,, \qquad (2.2.4.1)$$

where E_1 and E_2 are the error scores on, say, test 1 and test 2.

2.2.5 Assumption 5

Assumption five states that the error scores on one test (E_1) are uncorrelated with the true scores on another test (T_2). Mathematically, this assumption is expressed as

$$\rho_{E_1T_2} = 0 \,. \qquad (2.2.5.1)$$

2.2.6 Assumption 6

Assumption six states the definition of parallel tests. If X_a, T_a and σ_E^2 are observed score, true score and error variance of test *A* and X_b, T_b and $\sigma_{E'}^2$ are observed score, true score and error variance of test *B*, then test *A* and *B* are parallel tests when

$$\xi X_a = \xi X_b \text{ and } \sigma_E^2 = \sigma_{E'}^2 \,. \qquad (2.2.6.1)$$

2.2.7 Assumption 7

Assumption seven states that the tests that are essentially τ equivalent have true scores that are the same except for an additive constant, c_{12}.

2.3.0 Test Reliability

When simply defined, test reliability is a condition that fulfils the reproducibility of the test scores when the same test is administered again to the same population of examinees. In practice, it is difficult and rare to have a test with perfect reliability, i.e., a test which is capable of reproducing the same scores when administered again to the same population of examinees. A test which is highly reliable has its observed scores very close to true score. Therefore, technically, test reliability can be defined in terms of the reliability coefficient which is the squared correlation between the observed score and the true score of the test (Lord & Novick, 1968, p.61), meaning that the reliability reflects the observed score variance in terms of true score variance.

The test administrators always want to have a test with high reliability. A test with low reliability is a concern to the test administrators as it invites doubts on both consistency and utility of the scores obtained from the test.

Two broad sources of measurement errors have been classified to be responsible for non-reliability of a test. One of the categories of the error of measurement is called systematic errors of measurement. Algina & Crocker (1986, p.105) define systematic measurement errors as those errors which consistently affect an individual's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured. The other category of the error of measurement is called random errors of measurement. Algina & Crocker (1986, p.106) define random errors of measurement as purely chance happenings because of guessing, distractions in the test situation, administration errors, content sampling, scoring errors and fluctuations in the individual examinee's state.

It is clear from the two paragraphs that test reliability is dependent on the relationship between true scores, observed scores and errors of measurement. There are different ways of interpreting the reliability coefficient by involving true scores, observed scores and errors of measurement.

The procedures commonly used to estimate test score reliability are (1) alternate form method, (2) test-retest method, (3) test re-test with alternate forms and (4) split-half methods (Algina & Crocker 1986 & Yen & Allen,1979). The procedures are not described in the thesis.

2.3.1 Different ways of Interpreting the Reliability Coefficient of a Test

Yen & Allen (1979, p.73-75) give different ways of interpreting the reliability coefficient of a test in three different contexts.

Test Reliability in the Context of Parallel Tests: If a test X and a test X' are parallel tests, then the reliability coefficient of test X is the correlation of its observed scores with the observed scores of test X'. Mathematically, this can be written as

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2 \qquad (2.3.1.1)$$

Test Reliability in the Context of True Score and Observed Score:

(1) Reliability coefficient is the ratio of true score variance to observed score variance. Mathematically, this can be stated as

$$\rho_{X_1X_2} = \frac{\sigma_T^2}{\sigma_X^2}, \qquad (2.3.1.2)$$

where X_1 and X_2 are the observed scores of the test.

(2) Reliability coefficient is the square of the correlation between observed score and true score. Mathematically, this can be stated as

$$\rho_{X_1X_2} = \rho_{XT}^2 . \qquad (2.3.1.3)$$

Test Reliability in the Context of Observed Scores and Error scores:

(1) Reliability coefficient is one minus the squared correlation between observed and error scores. Mathematically, this can be stated as

$$\rho_{X_1X_2} = 1 - \rho_{XE}^2 . \qquad (2.3.1.4)$$

(2) Reliability coefficient is one minus the ratio of error score variance to observed score variance. Mathematically, this can be written as

$$\rho_{X_1X_2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} . \qquad (2.3.1.5)$$

2.3.2 Use of the Reliability Coefficient in Interpreting Test Scores

Yen and Allen (1979, p.76), offer a summary of applying the reliability coefficient in interpreting test scores. The summary is quoted in table 2.3.2.1.

Reliability Coefficient	Interpretation of s test scores
$\rho_{X_1X_2} = 1$	Measurement has been made without error
	Observed score is equal to true score for all examinees
	All observed-score variance reflects true-score variance
	All difference between observed scores reflect true-score differences
	The correlation between observed scores and true scores is 1
	The correlation between observed scores and error scores is 0
$\rho_{X_1X_2} = 0$	Only random error is included in the measurement
	Observed score is equal to error score for all examinees
	All observed variance reflect errors of measurement
	The correlation between observed scores and true scores is 0
	The correlation between observed scores and error scores is 1
$0 \le \rho_{X_1X_2} \le 1$	The measurement can include some error
	Observed score is equal to the sum of true score and error score
	Observed score variance includes some true score variance and some error variance
	Differences between scores can reflect errors of measurement as well as true score differences
	The correlation between observed scores and true scores equals the root of the reliability coefficient
	The correlation between observed scores and error scores is root of one minus the reliability coefficient
	Reliability is the proportion of observed score variance that is true score variance
	The larger the reliability coefficient is, the more confidently we can estimate true score from observed score, because error variance will be relatively smaller

Table 2.3.2.1: Different ways of interpreting test reliability coefficient.

2.4.0 Two popular Formulae for Estimating Reliability

In this section, two popular formulae for estimating test score reliability are briefly described. Coefficient alpha is used when parallel test forms are not available, where as the Spearman- Brown Formula is used when parallel test forms are available.

2.4.1 Internal-Consistency Reliability

Internal-consistency reliability is estimated using one test administration. A test is divided into two or more subtests, say, N subtests. The variances of scores of the subtests and the variance of the total test score are used to estimate the reliability of the test by using the formula (Yen & Allen1979, pp. 83-84) stated below:

$$\rho_{X_1X_2} \geq \alpha = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right], \qquad (2.4.1.1)$$

where X is the observed score for a test formed from combining N subtests, $X = \sum_{i=1}^{N} Y_i$, σ_X^2 is the population variance of X, $\sigma_{Y_i}^2$ is the population variance of the i[th] subtest, Y_i, and N is the number of subtests which combine to form X.

Equation 2.4.1.1 shows that coefficient Alpha is the lower bound of the test reliability, meaning that low Alpha does not provide good information about the actual reliability of a test.

Corollary:

1. If each subtest, Y_i, is a dichotomous item, equation 2.5.1.1 takes the following special form:

(a) $$\rho_{X_1X_2} \geq KR20 = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{N} p_i(1-p_i)}{\sigma_X^2}\right], \qquad (2.4.1.2)$$

where p_i is the proportion of examinees getting item i correct.

(b) $$\rho_{X_1X_2} \geq KR21 = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - N\bar{p}(1-\bar{p})}{\sigma_X^2}\right] \qquad (2.4.1.3)$$

where $\bar{p}$ is the average of the p-values.

2.4.2 The Spearman-Brown Formula

The Spearman-Brown formula expresses the reliability, $\rho_{X_1X_2}$, of a test in terms of the reliability, ρ_{YY_1} , of parallel subtests as

$$\rho_{X_1X_2} = \frac{N\rho_{YY_1}}{1+(N-1)\rho_{YY_1}}, \qquad (2.4.2.1)$$

where X=the observed score for a test formed from combining N subtests, $X = \sum_{i=1}^{N} Y_i$, Y_i is a subtest score that is a part of X, $\rho_{X_1X_2}$ is the population reliability of X, $\rho_{Y_1Y_2}$ is the population reliability of any Y_i , N is the number of parallel test scores that are combined to form X.

2.5.0 Standard Errors of Measurement and Confidence Intervals

When the discrepancy between an examinee's true score and observed score from a test is interpreted, confidence intervals are commonly used to show the interval in which the expected score is likely to fall. Standard error of measurement is used to calculate the confidence interval. The formula for the estimated standard error of measurement is

$$\sigma_E = \sigma_X \sqrt{1-\rho_{X_1X_2}}, \qquad (2.5.0.1)$$

where σ_X is the standard deviation for the observed scores for the entire examinee group and $\rho_{X_1X_2}$ is the test reliability estimate.

The confidence interval for an examinee's true score can be constructed as

$$x - z_c\sigma_E \le T \le x + z_c\sigma_E, \qquad (2.5.0.2)$$

where x is the observed score for the examinee, σ_E is estimated standard error of measurement, and z_c is the critical value of the standard normal deviate at the desired probability level.

A case from the Project Item Analysis for Booklet 11 Number of observations = 50 Number of items = 30 Results based on raw (unweighted) scores Mean = 22.480 S.D. = 4.535 Alpha = .788	The case information contains test reliability coefficient alpha estimated for a test consisting of 30 dichotomous items administered to 50 examinees. Alpha is 0.788 and standard deviation is 4.535. The observed scores of the examinees on this test consist of true scores and random errors. The magnitude of the random errors inherent in the test is 2.089. The mean of the true scores of the examinees may fall between 18.30 and 26.66 at 96 % confidence interval. On the whole, this test is a good test. The mean of the observed scores of the examinees is close to the mean of their true scores.

A case from the Project	The case information contains test reliability coefficient alpha estimated for a test consisting of 30 dichotomous items administered to 45 examinees.
Item Analysis for Booklet 12 Number of observations = 45 Number of items = 30 Results based on raw (unweighted) scores Mean = 23.822 S.D. = 3.702 Alpha = .711	Alpha is 0.711 and standard deviation is 3.702. The observed scores of the examinees on this test consist of true scores and random errors. The magnitude of the random errors inherent in the test is 2.00. The mean of the true scores of the examinees may fall between 19.82 and 27.82 at 96 % confidence interval. On the whole, this test is a good test. The mean of the observed scores of the examinees is close to the mean of their true scores.

2.6.0 Validity

Test scores are used for different purposes. For example, test scores are used for making placement decision, diagnosing learning difficulties, awarding grades, making admission decision, writing instructional guidance, setting future criterion, licensing and many more. In these examples, test scores provide scientific rationale for making inferences about examinees' behaviors in relation to their test scores. The test makers and the test users apply validity studies to make the inferences derived from the test scores useful for making decisions.

Glas et al., (2003, p.100) describe validity as the meaning, usefulness and correctness of the conclusions made from the test scores. Glas et al., (2003, p.100, cf. Messick, 1989, 1995) define validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment". Algina & Crocker (1986, p.217, cf. Cronbach, 1971) offer a procedural aspect of validity by emphasizing "validation as the process by which a test developer or test user gathers evidence to support the kinds of inferences that are to be drawn from test scores".

Lord and Novick (1968, p.61) define validity coefficient of measurement X with respect to a second measurement Y as the absolute value of the correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \qquad (2.6.0.1)$$

Implicit in the definitions of validity is the need to identify and describe the desired inferences that are to be drawn form the test scores before conducting the validation studies. The major types of validity are content validity, criterion validity and construct validity.

2.6.1 Content Validity

Content validity is the degree to which a test measures what it is intended to measure. Content validation study is engaged when a test user wants to make an inference from test scores to a larger domain of items similar to those used in the test. Content validity is concerned with sample population representativeness, meaning that the knowledge and skills contained in the test should be representative of the larger domain of knowledge and skills. Algina & Crocker (1986, p.217) have proposed the following steps for content validation study:

- *Defining the performance domain of interest*
- *Selecting a panel of qualified experts in the content domain*
- *Providing a structured framework for the process of matching items to the performance domain*
- *Collecting and summarizing the data from the matching process*

The proposed steps may be accompanied by a check list of questions presented in table 2.6.1.1 to assist preliminary planning tasks for the content validation study.

Q.No.	Question	Yes	No
1	Should domain objectives be weighted to reflect their importance?		
2	How the item- domain objective mapping task should be structured?		
3	What aspects of the item should be examined?		
4	How should content validation study result be summarized?		

Table 2.6.1.1: *Check list of questions for planning a content validation study*

2.6.2 Criterion Validity

Glas et al., (2003,p.101) define criterion validity as the extent to which the test scores are empirically related to criterion measures. Criterion-related validity exists in two forms known as (1) predictive validity and (2) concurrent validity. Predictive validity involves using test scores to predict criterion measurement that will be made at some point in the future, where as concurrent validity is the correlation between test scores and criterion measurement when both are obtained at the same time. Criterion-related validation study is used when a test user wants to make an inference from the examinee's test score to performance on some real behavioral variable of practical importance.

Algina & Crocker (1986, p.224) have proposed the following steps for criterion related validation study:

- *Identify a suitable criterion behavior and a method for measuring it.*
- *Identify an appropriate sample of examinees representative of those for whom the test will automatically be used.*
- *Administer the test and keep a record of each examinee's score.*
- *When the criterion data are available, obtain a measure of performance on the criterion for each examinee.*
- *Determine the strength of the relationship between test scores and criterion performance.*

Regression analysis can be applied to establish criterion validity. An independent variable could be used as a predictor variable, *X (Exam scores)*, and dependent variable, the criterion variable, *Y (Grade point averages).* The correlation coefficient between *X* and *Y* is called validity coefficient. It can be shown that the prediction of *Y* for the i^{th} person is

$$\hat{Y}_i = r_{XY}\left(\frac{S_Y}{S_X}\right)\left(X_i - \overline{X}\right) + \overline{Y}\,, \qquad (2.6.2.1)$$

where $\hat{Y}_i$ is the future grade point average, $\overline{Y}$ is the mean of the grade point averages, r_{XY} is the correlation coefficient of exam score *X* and grade point average *Y*, S_Y is the standard deviation of the grade point averages, S_X is the standard deviation of the exam scores, X_i is the exam score of i^{th} examinee.

This result can be expressed in confidence intervals by using the formula $\hat{Y}_i \pm z_c s_{Y.X}$, where z_c is the critical value from the normal table.

2.6.3 Reliability of Predictor and Criterion Validity

When reliability coefficient is expressed as

$$\rho_{X_1X_2} = \rho_{XT}^2 , \qquad (2.6.3.1)$$

it is clear that a test score cannot correlate more highly with any other variable than it can with its own true score. The maximum correlation between an observed test score and any other variable is

$$\sqrt{\rho_{XX'}} = \rho_{XT} . \qquad (2.6.3.2)$$

If test, *X*, is used to predict a criterion, *Y*, then ρ_{XY} is the validity coefficient. As ρ_{XY} cannot not be larger than ρ_{XT} , ρ_{XY} cannot be larger than $\sqrt{\rho_{XX'}}$, meaning that the square root of the reliability is the upper bound of the validity coefficient, $\rho_{XY} \leq \sqrt{\rho_{XX'}}$. Therefore, the reliability of the test affects the validity of the test.

If both a test score, *X*, and criterion score, *Y*, are unreliable, the validity coefficient, ρ_{XY} , may be attenuated relative to the value of the validity coefficient that would be obtained if *X* and *Y* did not contain measurement error. Yen & Allen (1979, p.98, cf. Spearman, 1904) present the correction for attenuation as below:

$$\rho_{T_XT_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} , \qquad (2.6.3.3)$$

where $\rho_{T_XT_Y}$ is the correlation between the true score for *X* and the true score for *Y*, ρ_{XY} is the correlation of observed scores *X* and *Y* containing error of measurement, $\rho_{XX'}$ is the reliability of observed score *X*, $\rho_{YY'}$ is the reliability of observed score *Y*.

Equation 2.6.3.3 expresses the correlation between true scores in terms of the correlation between observed scores and the reliability of each measurement. Lord and Novick (1968, p.70) interpret equation 2.7.3.3 as "giving the correlation between the psychological constructs being studied in terms of the observed correlation of the measure of these constructs and the reliabilities of these measures".

2.6.4 Construct Validity

Construct validity is the degree to which a test measures the theoretical construct or trait that it was designed to measure (Lord & Novick, 1968, p.278). Construct validation study is usually conducted by analyzing the observed score correlations of a test with another test based on the theory underlying the constructs being measured. If the theory of the constructs predict the two tests to correlate, then there should be appreciable correlation between the two tests for the tests to be valid, otherwise the tests do not measure the constructs. Yen & Allen (1979,pp.108-109) propose (1) group differences, (2) changes due to experimental interventions, (3) correlation and (4) process as the possible predictions which can be made during construct validation study, besides content and criterion validities.

Algina & Crocker (1986, p.230) have summarized the following steps as the general steps involved in conducting a construct validation study:

- *Formulate one or more hypotheses about how those who differ on the construct are expected to differ on demographic characteristics, performance criteria, or measures of other constructs whose relationship to performance criteria has been already validated. These*

hypotheses should be based on an explicitly stated theory that underlies the construct and provide its syntactic definition.

- *Select (or develop) a measurement instrument which consists of items representing behaviors that are specific, and concrete manifestations of the construct.*
- *Gather empirical data which will permit the hypothesized relationships to be tested.*
- *Determine if the data are consistent with the hypotheses and consider the extent to which the observed findings could be explained by rival theories or alternative explanations (and eliminate these if possible).*

2.7.0 Summary

CTT describes the relationship among observed score, true score and error of measurement. An observed score is the sum of true score and error of measurement. Expected value of the observed scores is the true score. Error score and true scores on a single test are uncorrelated. Error scores on two different tests are uncorrelated. Error scores on one test are uncorrelated with the true scores on another test. Two tests are parallel only when their corresponding true scores and error scores are equal. Essentially τ equivalent tests have same true scores and different additive constant.

Test reliability is the squared correlation between observed score and true score of the test. Two types of errors of measurement are systematic error of measurement and random error of measurement. For two parallel tests, the reliability coefficient is the correlation of the observed scores of one test with the observed scores of the other test. Reliability coefficient is the ratio of true score variance to observed score variance of a test. Reliability is the square of the correlation between observed score and true score. Reliability is the ratio of one minus the ratio of error score variance to observed score variance. Reliability is commonly estimated by using internal consistency reliability formula and Spearman-Brown Formula. Confidence interval is used to report true scores.

Validity is the absolute value of the correlation coefficient of two measurements. Three types of validity are content validity, criterion validity and construct validity. Criterion validity can be obtained by using regression analysis. Square root of the reliability is the upper bound of the validity.

2.8.0 References

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

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

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

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

Chapter 3

Item Parameters in CTT Context

3.0.0 Introduction

Item bank will be functional when it has large reserve of good items. The quality of the items in an item bank 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.

In this chapter the item parameters are discussed in the context of CTT.

3.1.0 Item Difficulty

Item difficulty (sometime known as item facility) is the proportion of examinees who answer an item correctly (Algina & 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. Mathematically, the definition based expression for p-value is

$$p_{ij} = \frac{\text{Number of persons with score i on item j}}{N}, \qquad (3.1.0.1)$$

where p_{ij} is the p-value for item *j* with score *i* and *N* is the total number of examinees who attempted the item *j*.

Corollary:

(1) For dichotomous items, p_j is equal to mean score of item *j*.

(2) For polytomous items,

$$p_j = \frac{(p_{ij})(X_{ij})}{\max_{all_i}(X_{ij})}, \text{ where } (X_{ij}) \text{ is the is score } i \text{ on item } j. \qquad (3.1.0.2)$$

(3) The mean of test score ($\overline{X}$) is

$$\overline{X} = \sum_{j=1}^{N} p_j, \qquad (3.1.0.3)$$

where N is number of items in the test.

Depending on the number of choices involved in dichotomous items, the p-values of items at which the maximum true score variance would be obtained also differ due to random guessing by those examinees who do not know the correct answer. Algina & Crocker (1986, p.313) provide an expression for the p-value of an item at which the true score variance would be maximum as

$$p_o = 0.50 + \frac{0.50}{m}, \qquad (3.1.0.4)$$

where p_o =observed p-value, and m =number of choices or alternatives or distractors.

The effect of random guessing is illustrated in table 3.1.0.1.

Number of Choices	Proportion Who Know Answer	Proportion Who Guess Answer	p_o	Lord's p_o
4-choice item	0.50	0.50/4	0.50+(0.50/4)=0.62	0.74
3-choice item	0.50	0.50/3	0.50+(0.50/2)=0.67	0.77
2-choice item	0.50	0.50/2	0.50+(0.50/2)=0.75	0.85

Table 3.1.0.1: Shows the effects of random guessing. Source: Algina & Crocker (1986, p.313).

From table 3.1.0.1, it is clear that an item difficulty increases with increase in number of alternatives when multiple-choice item format is used in the test. Lord's p_o is the demonstration made by Lord (Algina & Crocker, 1986, p. 313) which suggests that a test reliability can be improved by choosing items with *p*-values even higher than those computed by adjusting for random guessing.

3.1.1 Role of Item P-values in Item Analysis

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 miskeyed 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 distractor." Null distractors are indicative of obvious answers, nonparallel distractors, or nonsensical distractors.

3.2.0 Item Variance

Item variance is the square of the item standard deviation. Mathematically, item-variance can be expressed as

$$\sigma_j^2 = \frac{\sum \left(X_{ij} - \mu_j\right)^2}{N}, \qquad (3.2.0.1)$$

where X_{ij} = Score of examinee *i* on item *j*, μ_j = mean score on item *j*, and *N*= number of examinees.

Corollary:

For dichotomous items

(1) item variance can be calculated by using p-values as

$$\sigma_j^2 = p_j q_j, \qquad (3.2.0.2)$$

where q_j =1- p_j.

(2) standard deviation of the item can be calculated as

$$\sigma_j = \sqrt{p_j q_j}\ . \qquad (3.2.0.3)$$

3.2.1 Role of Item Variance in Item Analysis

Item variance indicates the variability of the answers to the item. A low item variance indicates that most students selected or presented the same response to the item (not necessarily the correct one). A high item variance means that a near even number of students selected or presented a particular response.

3.3.0 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.

In the following section, three commonly used item discrimination statistics are described.

3.3.1 Index of Item Discrimination

Index of item discrimination 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 (p_u) and lower group (p_l). The index of discrimination (*D*) is calculated as

$$D = p_u - p_l . \qquad (3.3.1.1)$$

3.3.1.1 Role of Index of Discrimination in Item Analysis

Algina & Crocker (1986, p.315,cf. 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.

3.3.2 Point Biserial Correlation

The point biserial correlation reflects the item-test correlation when a discrete binary variable (correct vs incorrect response to the item) is correlated with a continuous variable (total test score). Algina & Crocker (1986, p.317) state the mathematical sentence for point biserial correlation as

$$\rho_{pbis} = \frac{(\mu_+ - \mu_X)}{\sigma_X} \times \sqrt{\frac{p}{q}}, \qquad (3.3.2.1)$$

where μ_+ is the mean criterion score for those who answer the item correctly, μ_X is the mean criterion score for the entire group, σ_X is the standard deviation, *p* is the item difficulty, and *q* is the 1-*p*.

Corollary:

(1) Algina & Crocker (1986, p.324) present the minimum critical value of ρ_{pbis} that an item should have as 2 standard errors above the item criterion correlation (with minimum value equal to 0.00). The formula for estimating Pearson product moment standard error is given as

$$\hat{\sigma} = \frac{1}{\sqrt{N-1}}, \qquad (3.3.2.2)$$

where N is the sample size.

3.3.2.1 Role of Point Biserial in Item Analysis

A high (ρ_{pbis} > 0.4) point biserial correlation (ρ_{pbis} can range from -1 through 0 to +1) indicates that students answering the question correctly tended to do well on the test, while those students missing the item tended to be low scorers overall. This means that the item "discriminated" well between students knowing the exam material and those lacking such knowledge.

Correlation equal to zero suggests the item is either operating randomly or all the students gave the same response. Negative values should be carefully scrutinized, for the item is functioning in the opposite direction of the test (i.e. the higher the score, the less likely the student correctly answered the question).

3.3.3 Biserial Correlation Coefficient

Biserial correlation coefficient is usually used when the latent variable underlying item performance is assumed to be normally distributed. The correlation between the hypothetical latent trait variable underlying correct-incorrect response to test items and a continuously distributed criterion variable (total test score) is called Biserial correlation coefficient (Baker, 1997, p.11, cf. Lord & Novick,1968, p.337) . Algina & Crocker (1986, p.317, cf. Pearson, 1909) state the mathematical expression for the biserial correlation coefficient as

$$\rho_{bis} = \frac{(\mu_{+} - \mu_{X})}{\sigma_{X}} \times \left(\frac{p}{Y}\right), \qquad (3.3.3.1)$$

where μ_{+} is the criterion score mean of those who answered the item correctly, μ_{X} is the criterion score mean of all examinees, σ_{X} is the standard deviation, *p* is proportion of examinees who answered the item correctly, and Y is the ordinate of the standard normal curve at the z-score associated with p value for the item.

Corollary:

(1) The relationship between the biserial and Point biserial correlations is

$$\rho_{bis} = \frac{\sqrt{pq}}{Y} \times \rho_{pbis}. \qquad (3.3.3.2)$$

3.3.3.1 Role of Biserial Correlation Coefficient in Item Analysis

The role played by ρ_{bis} in item analysis is similar to the role played by ρ_{pbis}. However, Algina & Crocker (1986, p.318) note that the difference between ρ_{bis} and ρ_{pbis} is fairly moderate for items of medium difficulty, but as p-values drop below 0.25 or increase above 0.75, the difference increases sharply.

3.4.0 Item Reliability Index

Allen & Yen (1979, p.124) present the formula for item- reliability index (r_i) as

$$r_i = \sigma_i \rho_{pbis}, \qquad (3.4.0.1)$$

where σ_i and ρ_{pbis} are the standard deviation of item *i* and the point biserial correlation between the item score and the total test score.

3.4.0.1 Role of Item Reliability Index in Item Analysis

The role of item reliability index in making test is shown in equation 3.6.0.2. From equation 3.6.0.2, the contribution of each item to the magnitude of the standard deviation of a test score becomes clear to a test maker. Further, equation 3.6.0.3 shows that the test reliability is maximized when item reliability is maximized.

3.5.0 Item Validity Index

Allen, & Yen, (1979, p.124) gave the formula for item validity index as

$$v_{iy} = \sigma_i \rho_{pibsy}, \tag{3.5.0.1}$$

where σ_i and ρ_{pibsy} are the standard deviation of item *i* and the point biserial correlation between the item score and criterion score *y* respectively.

3.5.0.1 Role of Item Validity Index in Item Analysis

Equation 3.6.0.4 shows the role of item validity index. To achieve maximum validity of a test score, items with validity indices as large as their reliability indices have to be selected.

3.6.0 Making Test from Item Bank

As mentioned in the introduction, item bank is a reservoir of items. The items in the item bank have parameters discussed in sections 3.1.0 through section 3.5.0. The item parameters help a test maker to make a new test with maximum reliability and maximum criterion-related validity. Allen & Yen (1979, pp.124-125) present the following formulas for making test by using item bank.

$$\overline{X} = \sum_{i=1}^{k} p_i, \tag{3.6.0.1}$$

$$\widehat{S}_X = \sum_{i=1}^{k} r_i, \tag{3.6.0.2}$$

$$r_{XX'} = \frac{k}{k-1}\left[1 - \frac{\sum_{k=1}^{k} \sigma_i^2}{\left(\sum_{i=1}^{k} r_i\right)^2}\right], \tag{3.6.0.3}$$

$$\widehat{r}_{XY} = \frac{\sum_{i=1}^{k} v_{iY}}{\sum_{i=1}^{k} r_i} \tag{3.6.0.4}$$

where p_i is the p-value of item *i*, *X* is the score for the test composed of k items, $\overline{X}$ is the mean of test score *X*, $\widehat{S}_X$ is the estimated standard deviation of test score *X*, $r_{XX'}$ is the estimated reliability of test score *X*, $\widehat{r}_{XY}$ is the estimated validity of test score *X*, r_i is the item reliability index of item *i*, v_{iY} is the item validity index and *k* number of items selected from *N* number of items.

3.7.0 Summary

Item difficulty is the proportion of examinees who answer an item correctly. Item difficulty is also known by different names like p-value and item facility. Item with high p-values are easier than items with low p-values. When selecting items for a test, the items with p-values higher than 0.20 and lower than 0.95 and having positive point biserial should be selected.

Item variance is the square of item standard deviation. Items with high variance should be selected for making test.

Index of item discrimination is the difference between the p-values of examinees in upper group and lower group when groups are formed based on the cut score obtained by using the total test score.

Point biserial and biserial give information about how an item functions with respect to total test score. High point biserial correlation is indicative of an item being discriminative and vice versa.

Item reliability and item validity can be used while making test to enhance its reliability and validity. Maximum test reliability can be achieved when items with high point by serial correlations are used for the test. Test validity can be maximized by using items with validity indices as close to item reliability indices as possible.

A case from the Project

Item analysis for booklet 11

item	p-value	rit(u)	rit(w)	item	p-value	rit(u)	rit(w)
291	.960	.247	.203	305	.780	.344	.318
292	.920	.373	.349	306	.340	.380	.354
293	.800	.351	.376	307	.740	.485	.494
294	.860	.335	.312	358	.620	.219	.116
112	.960	-.113	-.132	308	.880	.433	.466
280	.880	.215	.200	314	.660	.607	.665
281	.840	.383	.347	315	.860	.043	.058
282	.760	.534	.550	316	.760	.163	.115
283	.760	.503	.498	317	.540	.390	.374
284	.760	.421	.419	318	.560	.618	.663
276	.780	.216	.159	323	.520	.684	.727
78	.980	.141	.095	324	.620	.392	.386
277	.840	.491	.537	325	.640	.456	.410
278	.920	.324	.363	326	.840	.371	.455
279	.500	.344	.250	327	.600	.464	.483

Number of observations = 50
Number of items = 30
Results based on raw (unweighted) scores
Mean = 22.480
S.D. = 4.535
Alpha = .788
Results based on weighted scores
Mean = 47.100
S.D. = 10.651
Alpha = .805
Corr(u,w) = .983

Analysis

Items 291, 112 and 78 have very high p-values. These items are very easy items. There answers might have been too obvious for the examinees. These items have to be eliminated or revised to make them useful as test items.

Item 112 has negative item-test correlation. The examinees who answered this item correctly might have scored less on the test as a whole. In other words, this item is likely to be incorrectly answered by the examinees who are likely to have high scores on the test. This item needs to be thoroughly scrutinized. The possible areas for studying this item are key answer, clue in the key answer and administrative directions.

Items 282, 283, 284, 277, 307, 314,318, 323, 325 and 327 have high item-test correlation. These items have functioned properly. They are attractive for the high scorers. These items are very discriminative.

Item 315 has item-test correlation of 0.043 which is almost zero. This item has unpredictable behavior. This item is likely to be attractive to both high scorers and low scorers. The item needs to be checked.

On the whole, items appear to be moderately easy and not very discriminative.

A case from the Project

Item analysis for booklet 12

item	p-value	rit(u)	rit(w)	item	p-value	rit(u)	rit(w)
323	.756	.364	.386	330	1.000	9.99	9.99
324	.711	.327	.342	331	.978	.074	.045
325	.711	.420	.449	332	.867	.158	.174
326	.911	.533	.599	333	.600	.365	.301
327	.911	.259	.275	334	.867	.228	.190
314	.867	.511	.573	342	.667	.552	.604
315	.867	.493	.469	343	.933	.083	.070
316	.733	.405	.353	344	.822	.386	.424
317	.800	.291	.279	345	.667	.386	.374
318	.800	.426	.456	346	.911	.344	.294
305	.867	.299	.302	353	.667	.195	.178
306	.333	.340	.328	354	.756	.545	.563
307	.844	.211	.189	355	.533	.460	.450
279	.667	.183	.104	356	.911	.090	.075
308	.978	.074	.045	357	.889	.403	.383

Number of observations = 45
Number of items = 30
Results based on raw (unweighted) scores
Mean = 23.822
S.D. = 3.702
Alpha = .711
Results based on weighted scores
Mean = 51.467
S.D. = 8.191
Alpha = .724
Corr(u,w) = .987

Analysis

RIT(U) stands for unweighted item test correlation and RIT(W) stands for weighted item test correlation. The weight assigned to an item is equal to its discrimination value.

Item 330 has p-value of 1. This item can be eliminated without further study. Its item-test correlation is out of range (9.99)!

Items 308, 331 and 343 have high p-values. The item-test correlations of these items are almost zero. They are either functioning unpredictably or almost all the examinees got them correct. These items may have ceiling effects on the proficient examinees. These items need revision.

Item 306 has low p-value, meaning this item is a difficult item. Its item test-correlation is 0.33 which is indicative of it being quite discriminative.

Item 356 has high p-value. Its item-test correlation is 0.090 which is indicative of it being unpredictable. This item warrants revision.

3.8.0 Reference

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

Baker, R. (1997). *Classical Test Theory and Item Response Theory in Test Analysis*: Special Report No. 2: Language Testing Update, SDS Supplies Limited, Preston Lancashire.

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

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

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

Chapter 4

Making Norm Reference Table

4.0.0 Introduction

Broadly speaking, a test can be categorized into two categories known as criterion referenced test and norm referenced test. Depending on which of these two categories a test fits in, the ways of interpreting the test scores will be different. For a criterion referenced test, test raw scores can sufficiently describe the examinees' strengths and weakness in the construct of interest. However, for a norm referenced test, raw scores will not have meaningful interpretations about the examinees' strengths and weakness in the construct of interest without the comparative information about the normative population (Crocker & Algina,1986, p.431). Norms are commonly used to enhance the interpretability of the raw scores of norm referenced test.

The Reading and the Vocabulary tests described in chapter 1 are norm referenced tests and as such a list of steps for conducting norming study necessary for making norm scale followed by a brief description of making norm reference table or norm scale will be presented in this chapter.

4.1.0 Norming Study

Norming study is a way of choosing sample population and designing appropriate probability sampling methods to collect the required information about the population of interest. Crocker & Algina (1986, p.432) provide the following steps for conducting norming study:

1. *Identify the population of interest (e.g., all students in a particular school district or all applicants for admission to a particular program of study or type of employment).*

2. *Identify the most critical statistics that will be computed for the sample data (e.g., mean, standard deviation, percentile ranks).*

3. *Decide on the tolerable amount of sampling error (discrepancy between the sample estimate and the population parameter) for one or more of the statistics identified in step 2. (Frequently the sampling error of the mean is specified).*

4. *Devise a procedure for drawing a sample.* Some commonly used procedures for drawing a sample are (1) simple random sampling, (2) systematic sampling and (3) stratified random sampling.

5. *Estimate the minimum sample size required to hold the sampling error within the specified limit. Various formulas must be used depending on the sampling strategy employed.*

6. *Draw the sample and collect the data. Document the reason for any attrition which may occur. If substantial attrition occurs (e.g., failure of an entire school to participate after it has been selected into the sample), it may be necessary to replace this unit with another chosen by the same sampling procedure.*

7. *Compute the values of the group statistics of interest and their errors.*

8. *Identify the types of normative scores that will be needed and prepare the normative score conversion tables.*

9. *Prepare written documentation of the norming procedure and guidelines for interpretation of the normative scores.*

The details of these steps are beyond the scope of this chapter and no further elaborations will be made on them .However, they are mentioned in the chapter as they make up the procedures prior to making norm table or norm scale.

4.2.0 Making Norm Table

Norm table consists of normative scores which will enable the raw scores to be interpreted in terms of their relative locations and frequencies within the total score distribution (Crocker & Algina, 1986, p.439). The commonly used normative scores are percentile rank, normalized z-scores, Stanines, scaled scores and age and grade equivalent scores. In the following sections, percentile rank is described, as it is used in making test norms in the project.

4.3.0 Percentile Rank

Allen & Yen, (1979, p.150) define percentile rank of a trait value as the percentage of people in a norm group who have trait values less than or equal to that particular trait value. Crocker & Algina, (1986, p.439) present the mathematical sentence for the definition of percentile rank as

$$P = \frac{cf_l + 0.5(f_i)}{N} \times 100\%, \qquad (4.3.0.1)$$

where cf_l is the cumulative frequency for all scores lower than the score of interest, f_i is the frequency of scores in the interval of interest, and *N* is the number of the sample.

The following steps are implicit in equation 4.3.0.1:

a. Construct frequency distribution for the raw scores.

b. For given raw score, determine the cumulative frequency for all scores lower than the score of interest.

c. Add half the frequency of the score of interest to the cumulative frequency value determined in step 2.

d. Divide the total by *N*, the number of examinees in the norm group and multiply by 100%.

4.4.0 Summary

Two broad categories of test are criterion referenced test and norm referenced test.

The raw scores of the criterion referenced test are sufficient to describe the proficiency of examinees in the construct of interest.

The raw scores of the norm referenced test will have meaningful interpretation when they are presented with norm scores. Norm table consists of normative scores which will enable the raw scores to be interpreted in terms of their relative locations and frequencies within the total score distribution.

Different normative scores are percentile rank, normalized z-scores, Stanines, scaled scores and age and grade equivalent scores.

Percentile rank of a trait value is the percentage of people in a norm group who have trait values less than or equal to that particular trait value.

A case from the Project

A copy of the distribution obtained for Booklet 1 for Raw Scores

Norm Table

Score Distribution

Score	KB/GL	BB+/BB	HA/VW
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	1	0	0
6	1	0	0
7	2	1	0
8	2	1	0
9	3	1	0
10	5	2	0
11	6	2	0
12	8	3	1
13	10	4	1
14	12	5	1
15	15	7	1
16	18	9	2
17	21	11	2
18	25	13	3
19	29	16	4
20	34	20	6
21	40	24	7
22	46	29	10
23	52	35	13
24	59	42	17
25	67	50	23
26	75	60	31
27	83	71	42
28	91	82	58
29	97	93	80
30	100	100	100

The case information shows the distribution of raw scores in terms of percentile rank.

4.5.0 Reference

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

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

Chapter 5

Item Response Theory

5.0.0 Introduction

In the previous chapters, classical test theory and its functions in item banking and test construction were described. The item parameters and reliability estimates obtained by using classical test theory were shown to depend on examinee sample. The expected scores of examinees were also shown to depend on the types of test items. Finally, classical test theory measures the performance of an examinee at test level instead of measuring the same at the item level. These are some limitations of the classical test theory. However, an alternative field of study, popularly known as item response theory (IRT), has been pursued by the measurement specialists and the psychometricians alike and succeeded in finding solutions to the constraints posed by the classical test theory.

IRT is used in the project to make item bank for reading and vocabulary tests and to construct tests on them by using the item bank.

In this chapter IRT will be introduced with its underlying assumptions.

5.1.0 Item Response Theory

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 & 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.

5.2.0 Assumptions of Item Response Theory

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

5.2.1 Dimensionality of Latent Space

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 & 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 & 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.

5.2.2 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 & 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.

Hambleton, Swaminathan, & Rogers (1991, p.10) present the mathematical definition of local independence as

$$P(U_1,U_2,...,U_n \mid \theta) = P(U_1 \mid \theta)P(U_2 \mid \theta)...P(U_n \mid \theta)$$

For $P_i(\theta) = P(U_i = 1 \mid \theta)$ and $Q_i(\theta) = P(U_i = 0 \mid \theta)$

$$\begin{aligned} &P(U_1,U_2,...,U_n \mid \theta) \\ &\quad = P(U_1 \mid \theta)P(U_2 \mid \theta)...P(U_n \mid \theta) \\ &\quad = P_1(\theta)^{u_1} Q_1(\theta)^{1-u_1} P_2(\theta)^{u_2} Q_2(\theta)^{1-u_2} ...P_n(\theta)^{u_n} Q_n(\theta)^{1-u_n} \\ &\quad = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} , \end{aligned} \qquad (5.2.2.1)$$

where θ is the ability assumed to influence the ability of the examinee on the test, U_i is the response of a randomly chosen examinee to item *i* (*i*=1,2,…,*n*) and $P(U_i = 1 \mid \theta)$ is the probability of a correct response, and $P(U_i = 0 \mid \theta)$ is the probability of an incorrect response.

According to equation 5.2.2.1, the local independence is the probability of a response pattern on a set of items which is equal to the product of probabilities associated with the examinee's responses to the individual items.

5.2.3 Item Characteristic Curves

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 & Swaminathan, 1985, p.25). For an item *i* with binary scores (0 and 1), the item characteristic curve is defined by the expression

$$P_i(U_i \mid \theta) = P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} , \qquad (5.2.3.1)$$

where $P_i(U_i \mid \theta)$ is equal to $P_i(\theta)$ when u_i=1 and $P_i(U_i \mid \theta)$ is equal to $Q_i(\theta)$ when u_i=0.

Lord & 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 & 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 & Swaminathan, 1985, p.18).

5.2.4 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 & 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.

5.3.0 Summary

Item response theory states that examinee test performance can be predicted by a set of factors called traits and the relationship between examinee test performance and traits assumed to be influencing item performance can be described by monotonically increasing function called item characteristic curve. Five common assumptions of Item Response Theory are (a) dimensionality of latent space, (b) local independence, (c) item characteristic curves (d) item and ability parameters invariance and (e) speededness.

5.4.0 References

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

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

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

Chapter 6

One-Parameter Logistic Model (OPLM)

6.0.0 Introduction

The OPLM is an item response model which combines the attractive mathematical properties of the Rasch model with the flexibility of the two parameter logistic model (Verhelst et al., 1995). This means that the OPLM has the property of specific objectivity (Baker, 1992, pp. 134-136) and sufficient statistic of raw score for the ability parameter (Fischer, 1995, pp. 15-25) which are the main strengths of the Rasch model and has the flexibility of the two parameter logistic model when discrimination indices are imputed as known integer constants. The OPLM is applicable to both dichotomous response data and polytomous response data.

The conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation procedures are used to estimate the item parameters.

6.1.0 Presentation of One Parameter Logistic Model

As formulated by Verhelst et al., (1995), if the response to item *i*, denoted by X_i, falls in the score range (0,1,…,m_i), the probability of observing $X_i = j$ as a function of ability parameter θ is given by

$$\psi_{ij}(\theta) = \Pr(X_i = j \mid \theta) = \frac{\exp\left[a_i\left[j\theta - \sum_{g=1}^{j}\beta_{ig}\right]\right]}{1+\sum_{h=1}^{m_j}\exp\left[a_i\left[h\theta - \sum_{g=1}^{h}\beta_{ig}\right]\right]}, \qquad (j=0,\ldots,m_i), \qquad (6.1.0.1)$$

where $\beta_{ig}, g = 1,\ldots, m_i$ are the parameters of item *i*, and a_i stands for a known discrimination index for item *i* . The function $\psi_{ij}(\theta)$ is called the item category response function.

According to Verhelst et al., (1995), the following aspects are embedded in equation 6.1.0.1:

(a) If it is assumed that $S = \sum_{i=1}^{k} a_i X_i$ is a sufficient static for the ability parameter, equation 6.1.0.1 holds.
(b) If the discrimination indices are integer constants, item parameters can be estimated using CML procedures.
(c) The structure of the equation is such that the testing procedures can be devised with a focus on the validity of the selected discrimination indices and are informative with respect to the direction in which the information must be changed to obtain model fit.

6.2.0 Goodness of Fit Test for the Model

The overall test of model fit for the OPLM are constructed in such a way that they have power against specific model violations (Verhelst et al., 1995). The tests focus on two important alternative hypotheses which are (a) incorrect specification of the discrimination indices and (b) differences in item functioning in different groups (DIF). The OPLM generates a family of generalized person tests introduced by Glas and Verhelst (1989, 1995) as stated in Verhelst et al., (1995).The different tests are (a) M_i statistics, (b) S_i and (c) R statistics.

6.2.1 The M_i Tests

For the OPLM, the person's weighted sum score *s* is a sufficient statistic for the latent ability θ and the probability of obtaining a correct response given *s*, denoted by $\pi_i \mid s$ will be a fair approximation of the item response function obtained in equation 6.1.0.1 and so will its CML estimate $\hat{\pi}_i \mid s$ (Verhelst, et al., 1995). Generally the graph of $\hat{\pi}_i \mid s$ is S-shaped and dependent on the discrimination index a_i. However, if a_i has too large a value, one can expect atypical pattern of deviations of the observed proportions $p_i \mid s$ from their predicted values $\hat{\pi}_i \mid s$. For small *s*, $p_i \mid s > \hat{\pi}_i \mid s$ and for large *s*, ($p_i \mid s$)<($\hat{\pi}_i \mid s$). Accordingly, if the scores are portioned in a low group (L), a medium group (O), and a high group (H), the *M* static is formulated as

$$M_i^* = \sum_{s \in L} \left[(Pi \mid s) - (\hat{\pi}_i \mid s) \right] - \sum_{s \in H} \left[(Pi \mid s) - (\hat{\pi}_i \mid s) \right]. \qquad (6.2.1.2)$$

The statistic M_i^* will tend to positive if the discrimination index is set too high and negative if discrimination index is set too low. The positive M_i^* statistic suggests downward adaptation of discrimination index and the negative M_i^* statistic suggests upward adaptation of discrimination.

Tip

If $M_i^* < -2$ suggests an upward adaptation of a_i and $M_i^* > 2$ suggests a downward adaptation of a_i.

Hemker, B.T

The OPLM generates three versions of the M_i^* test based on the score partition into a low, a medium, and a high group. According to Verhelst et al., (1995), the M_i^* tests are designed as follows.

> *First, the scores on the booklets in which the focused item i appears are ordered in such a way that the conditional probabilities, $\pi_{i|sb}$, of obtaining a correct response on item i given the sum score on booklet b are monotonously increasing. So, each score s receives a rank number w(s,b) such that $\pi_{i|sb} < \pi_{i|s'b'} \Rightarrow w(s,b) < w(s',b')$, where s and s' are scores obtained from the same or different booklets b and b' .In the event of ties, the ordering within ties is random. Once the scores are ordered, three definitions of the low, medium, and high score groups can be applied as M_i , $M2_i$, and $M3_i$ respectively.*
>
> *For the M_i test, the low score group is composed of scores having a low (at most 0.4) conditional probability of having item i correct, and the high score group has a probability of correct of at least 0.6, resulting in $L(M_i) = \{s \mid \pi_{i|sb} \leq 0.4\}$ and $H(M_i) = \{s \mid \pi_{i|sb} \geq 0.6\}$.*
>
> *For the $M2_i$ tests, two groups are formed: the low and high score group contain approximately half of the sub sample which has responded to item i . The approximation is caused by the fact that individuals having the same booklet-score combination are always allocated to the same group.*

For the $M3_i$ tests, three groups are formed, the score groups contain approximately 33% of the observations and are associated with a three-way partition of the range of rank numbers.

In the table 6.2.1.1, a section of the OPLM output obtained for the analysis of a vocabulary test is reproduced for illustration of the M_i statistics.

Nr	Label	A	B	SE(B)	S	DF	P	M	M2	M3
107	80134W	1	-.561	.129	33.426	6	.000	0.933	-3.040	-3.167
16	74017W	5	-.152	.065	6.207	1	.013	-.950	.544	.166
115	74007W	1	-.201	.118	29.655	7	.000	-.061	3.913	3.779

Table 6.2.1.1: Shows M_i statistics.

In the table, M2 and M3 for item 107 indicate the need to increase the discrimination parameter (A) for medium and high score groups. Item 16 fits the model well. M2 and M3 for item 115 indicate the need to decrease A for medium and high score groups, but A is already at the lowest value. For further details on M tests, readers may like to refer Fischer & Molenaar, (eds., 1995, pp.70-95; 216-237; 326-351).

6.2.2 The S_i Tests

The S_i tests is based on the differences between the observed and expected proportion of responses in homogeneous score groups and the statistics provide a test for the fit of specific item and have power against misspecifications of the discrimination indices. S_i statistics is also used to detect differential item functioning.

The statistic is defined as follows (Verhelst et al.,1995):

The scores from different booklets are equated by assigning rank numbers w(s,b). The range of the rank numbers is partitioned into equivalence classes $G_q, q = 1,...,Q$. For the partitioning of the score range, two restrictions are introduced:

$$\sum_{w(s,b)\in G_q} n_{sb} \geq 30, (q = 1,...,Q), \text{ and for each dichotomization [:j+1], (j=0,…, } m_{i-1}),$$

$$\min\left[\sum_{w(s,b)\in G_q} n_{sb} \sum_{h=0}^{j} \hat{\pi}_{ih|sb}, \sum_{w(s,b)\in G_q} n_{sb} \sum_{h=j+1}^{m_i} \hat{\pi}_{ih|sb} \right] \geq 5, \text{ for } q = 1,...,Q.$$

Let the random variable $M_{ij|sb}$, with realization $m_{ij|sb}$, be the number of responses in category j of item i for respondents with sum score s of booklet b, and let $d_{ij|sb}$ be the difference between $m_{ij|sb}$ and its expected value, that is,

$$d_{ij|sb} = m_{ij|sb} - n_{sb}\hat{\pi}_{ij|sb}. \qquad (6.2.1.3)$$

For a dichotomization [:j+1], a vector d_i has entries $d_{iq}, (q = 1,...,Q)$, defined by

$$d_{iq} = \sum_{w(s,b)\in G_q} \sum_{h=j+1}^{m_i} d_{ih|sb}, \quad (q = 1,...,Q)$$

The statistic S_i is then defined as

$$S_i = d_i' V^{-} d_i . \qquad (6.2.1.4)$$

S_i has an asymptotic chi-square distribution with Q-1 degrees of freedom, where V_i^{-} is a generalized inverse of the estimated asymptotic covariance matrix of d_i.

For further details on S_i tests, readers may like to refer Fischer & Molenaar (eds., 1995, pp.70-95; 216-237; 326-351).

In the table 6.2.2.1, a section of the OPLM output obtained for the analysis of a vocabulary test is reproduced for illustration of the S_i statistics.

Nr	Label	A	B	SE(B)	S	DF	P	M	M2	M3
1	75002W	1	-.315	.158	4.448	6	.616	-.413	-.262	.528
16	74017W	5	-.152	.065	6.207	1	.013	-.950	.544	.166
107	80134W	1	-.561	.129	33.426	6	.000	.933	-3.040	-3.167

Table 6.2.2.1: Shows S_i statistics.

From the table, it can be concluded that item 1 (S=4.448; df=6; p=.616) fits the model, item 16 (S= 6.207;df=1;p=.013) is just around the critical value of the fit statistic, depending on the critical value for the fit statistic (Fischer & Molenaar ,Eds., 1995,pp.235), and item 107 (S=33.426;df=6;p=.000) does not fit the model.

The OPLM provides summary of S-tests with distribution of p-values as shown below:

0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. ← p-values
0/ 12/ 7 11 18 14 13 11 14 10 11 17 ← Items

Fig. 6.2.2.1: Distribution of p-values for S-tests when items fit the model.

The items falling in the first and the second slashes indicated as 0.--/---/---.1 have p-values equal to or less than 0.01 and 0.05 respectively. In fig. 6.2.2.1, there are 0 items with p-values equal to or less than 0.01 and 12 items with p-values equal to or less than 0.05.This is a case of a good fit. The distribution of items is almost rectangular or uniform.

The next illustration for the distribution of p-values for S-tests is a case of a bad fit. The distribution of items is skewed.

0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1.
23/ 18/ 13 26 20 11 9 7 6 8 6 7

Fig.6.2.2.2: Distribution of p-values for S-tests when items do not fit the model.

In fig. 6.2.2.2, there are 23 items whose p-values are less than or equal to 0.01 and 18 items whose p-values are equal to or less than 0.05.

6.2.3 The R_{1c} Test

R_{1c} is the global test for the goodness of fit between the model and the data. It shows how far the data meets one of the assumptions of the model like monotone increasing item characteristic curves.

The R_{1c} test is defined as follows (Verhelst et al.,1995):

The score continuum of a booklet b is partitioned into Q_b (maximal four) equivalence classes G_{bq}, q =1,…,Q_b. The vector d_{bq} has elements $d_{bq}(i,j),(j=1,...,m_i; i \in I_b)$, defined by

$$d_{bq}(i,j) = \sum_{w(s,b)\in G_{bq}} d_{ij|sb}. \qquad (6.2.1.5)$$

The number of elements in d_{bq} is $Q_b \sum_{i\in I_b} m_i$. The statistic R_{1c} is defined as

$$R_{1c} = \sum_b \sum_q d'_{bq} W^-_{bq} d_{bq}, \qquad (6.2.1.6)$$

where W_{bq} is a generalized inverse of the estimated covariance matrix of d_{bq}, has an asymptotic chi-square distribution with degrees of freedom given by

$$df(R_{1c}) = \sum_b Q_b \sum_{i\in I_b} m_i - \sum_b Q_b - (B-1) - \sum_i^k m_i \ . \qquad (6.2.1.7)$$

The equation(6.2.1.6) *is based on the rationale that since the person's sum score is a sufficient statistic for ability, checking $\sum_{W(s,b)\in G_q} m_{ij|sb}$ against its expected value across various levels may reveal differences between the observed and expected frequency of responding in a certain category at various ability levels. If, for instance, for dichotomous items , the observed number of positive responses is too small at low score levels and too large at high score levels, the item characteristic curve is steeper than predicted by the model, that is, the discrimination index selected is too small. For polytomous items, the discrimination index that is too small will result in too few responses in the middle category and too many responses in the extreme categories. To interpret the magnitude of the differences, scaled deviates can be used. These are obtained by dividing the differences $\sum_{w(s,b)\in G_q} d_{ij|sb}$ by their standard error, thus producing a standardized binomial variable. Squaring and summing across categories and equivalence classes for the same item results in an indication of the contribution of the item to the fit statistic.*

For further details on R_{1c} , readers may like to refer Fischer & Molenaar (Eds., 1995,pp.325-351).

The summary of R1c statistic obtained for a reading test is reproduced below for illustration purpose.

SUMMARY OF R1c-STATISTICS

booklet	#items	#groups	#deviates	sum sq.	R1c
1	29	3	84	110.23	
2	29	2	56	61.55	
3	29	3	84	88.90	
4	29	4	112	112.44	
5	28	4	108	118.38	

R1c* = 1115.710; df = 773; p = .0000
The reported R1c*-statistic is an approximation to R1c.

Tip

Although the R1c statistic here indicates that the model does not fit the data, in practice the ratio of R1c to df less than 1.5 is a good indication of the model being close to fit. This may be acceptable for practical purpose.

Hemker, B.T

The tip in the box implies the implication of the power of the test when sample size is very large.

R1c-components of item categories

Item Nr.	label	cat	sum of sq.	#terms	average	#dev>2
307	40042_ Le	1	.003	2	.001	0
355	50045_Le	1	.007	2	.003	0
343	04038_Le	1	.014	2	.007	0
253	40070_Le	1	12.152	4	3.038	1
160	30133_Le	1	27.542	9	3.060	3
172	50052_Le	1	30.359	9	3.373	1
197	50081_Le	1	32.213	9	3.579	4
317	30081_Le	1	7.501	2	3.750	1
323	50067_Le	1	14.306	2	7.153	2

Tip

In the illustration, item 307 (average=0.001) does not contribute to the misfit of the model unlike the item 323 (average=7.153) which contributes so much. There is no definite guideline for the interpretation. If the average is greater than 3, it indicates the need to study how the item functions across booklets. If R1c shows no fit (p=0.000), it is advisable to look at the items with highest sum and items with average greater than or equal to 3 as they are suspects of misfit.

Hemker, B.T

6.3.0 Parameter Estimation

The OPLM uses conditional maximum likelihood and marginal maximum likelihood estimation methods to estimate item parameters and person parameters. In this section the main equations for these two estimation methods will be given. For details, the readers may like to refer (a) Emretson & Reise (2000, pp.210-218), (b) Fischer & Molenaar (Eds., 1995, pp.44-49; 219-224), (c) Baker (1992, pp.136-144) and (d) Hambleton & Swaminathan (1985, pp.138-141).

6.3.1 Conditional Maximum Likelihood Estimation (CML)

CML estimates the item parameters by equating the sufficient statistics with their expected value conditionally on the frequency distribution of the persons' sum score (Verhelst et at., 1995, p.9-10).The CML equation for item parameter estimation of OPLM is

$$t_{ij} = E\left(T_{ij} \mid n, \beta\right), \quad (j = i,1,...,m_i; i = 1,...,k) \qquad (6.3.1.1)$$

In equation 6.3.1.1, β denotes the vector of the item parameters, *n* denotes the vector of elements n_s for *s*=0,…, S_b, where n_s denotes the number of persons obtaining score *s* for *s*=0,…, S_b and S_b is the weighted sum score on booklet *b*, i.e. $S_b = \sum m_i a_i$ and T_{ij} denotes the counts of the number of responses in category *j* for every item *i* for *j*=1,..., m_i, with realization t_{ij}.

6.3.2 Marginal Maximum Likelihood Estimation (MML)

OPLM uses MML estimation method to estimate population parameters. By assuming that θ has a normal distribution with mean μ and standard deviation σ, (Verhelst et al., 1995, p.10-11) present the MML equations as

$$t_{ij} = \sum_s n_s E(\psi_{ij}(\theta) \mid s, \beta, \mu, \sigma), \qquad (j = 1,...,m_i; i = 1,...,k), \tag{6.3.2.1}$$

$$\mu = n^{-1} \sum_s n_s E(\theta \mid s, \beta, \mu, \sigma), \tag{6.3.2.2}$$

$$\sigma^2 = n^{-1} \sum_s n_s E(\theta^2 \mid s, \beta, \mu, \sigma) - \mu^2, \tag{6.3.2.3}$$

In equation 6.3.2.1, $\psi_{ij}(\theta)$ has the meaning as defined in equation 6.1.0.1, *n* denotes the sample size and the other symbols have the meaning as defined in equation 6.3.1.1.

6.4.0 OPLM and other Models

OPLM has many properties similar to Rasch model, two-parameter logistic model and generalized partial credit model.

$$\psi_{ij}(\theta) = \Pr(X_i = j \mid \theta) = \frac{\exp\left[a_i \left[j\theta - \sum_{g=1}^{j} \beta_{ig} \right] \right]}{1 + \sum_{h=1}^{m_j} \exp\left[a_i \left[h\theta - \sum_{g=1}^{h} \beta_{ig} \right] \right]}, \ (j=0,\ldots,m_i) \tag{6.4.0.1}$$

Upon close examination of OPLM as presented in equation 6.4.0.1, a number of observations are noticeable.

6.4.1 OPLM and Rasch Model

When response category *m* is one and item discrimination parameter *a* is one, OPLM takes the form of Rasch model (Hambleton et al., 1991, pp. 12-14) as

$$\psi_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}, \quad \text{where } i = 1,2,...,n. \tag{6.4.1.1}$$

6.4.2 OPLM and Two-Parameter Logistic Model

When response category *m* is one and item discrimination parameter *a* is not uniform but taken as the known integer constants by imputing integer values, OPLM takes the form of two parameter logistic model ((Hambleton et al., 1991, pp.14-16) as

$$\psi_i(\theta) = \frac{\exp a_i(\theta - \beta_i)}{1 + \exp a_i(\theta - \beta_i)}, \ \text{where } i = 1,2,...,n. \tag{6.4.2.1}$$

However, unlike in OPLM, the values of item discrimination parameters in two parameter logistic model are estimated.

6.4.3 OPLM and Partial Credit Model

When response category *m* is more than one and item discrimination is one, OPLM takes the form of partial credit model (Masters & Wright 1997, pp.101-105) as

$$\psi_{ij}(\theta) = \Pr(X_i = j \mid \theta) = \frac{\exp\left[j\theta - \sum_{g=1}^{j} \beta_{ig} \right]}{1 + \sum_{h=1}^{m_j} \exp\left[h\theta - \sum_{g=1}^{h} \beta_{ig} \right]}, \ (j=0,\ldots,m_i). \qquad (6.4.3.1)$$

6.4.4 OPLM and Generalized Partial Credit Model

When response category *m* is more than one and item discrimination is not uniform but taken as the known integer constants by imputing integer values, OPLM takes the form of generalized partial credit model (Muraki, 1997, pp.153-156) as

$$\psi_{ij}(\theta) = \Pr(X_i = j \mid \theta) = \frac{\exp\left[a_i \left[j\theta - \sum_{g=1}^{j} \beta_{ig} \right] \right]}{1 + \sum_{h=1}^{m_j} \exp\left[a_i \left[h\theta - \sum_{g=1}^{h} \beta_{ig} \right] \right]}, (j=0,\ldots,m_i). \qquad (6.4.4.1)$$

However, unlike in OPLM, the values of item discrimination parameters in generalized partial credit model are estimated.

6.5.0 Strengths of OPLM

It is shown how OPLM resembles other models. OPLM is known to exhibit all properties of Rasch model like mono-tone item characteristic function with increase in latent trait, unidimensionality, local independence, sufficient statistic and specific objectivity.

The existence of sufficient statistic warrants the use of CML estimation method for estimation of item parameters.

OPLM has the flexibility of being both strong model and weak model (Hemker, 1996, chapter 1, p.3), meaning it can be used for the data that require either a strong model or a weak model.

By taking item discrimination parameter as a known constant integer and imputing it in the model, OPLM exhibits the properties of other models. The precision of the values of the imputed item discrimination parameters are measurable with statistics. The S-statistic shows whether an item discrimination parameter of an item is precisely specified or mis-specified. In case of misspecification, there are M-statistics which dictate addition or subtraction of integer number from the mis-specified item discrimination parameter.

OPLM has R-statistic which gives the global goodness of fit test between the model and the data. It also indicates which item contributes how much to the misfit between the model and the data, meaning that misfitting items are detectable at global level as well.

6.6.0 Summary

The OPLM is an item response model which combines the attractive mathematical properties of the Rasch model with the flexibility of the two parameter logistic model.

The S_i statistics provide a test for the fit of specific item and have power against misspecifications of the discrimination indices. The statistic M_i^* will tend to be positive if the discrimination index is set too high and negative if discrimination index is set too low. R_{1c} is the global test for the goodness of fit between the model and the data.

The conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation procedures are used to estimate the item parameters.

OPLM can take the forms of other models like two logistic parameter model, partial credit model and generalized partial credit model depending on item discrimination parameter and response data.

6.7.0 References

Baker, B.F.(1992, pp.134-136). *Item Response Theory, Parameter Estimation Techniques*: Marcel Dekker, Inc.

Emretson, E. S. & Reise, P.S. (2000, pp.210-218). *Item Response Theory for Psychologists*: Lawrence Erlbaum Associates, Publishers.

Fischer, H.G. (1995, pp. 15-38). Derivations of the Rasch Models. In In Fischer, H.G. & Molenar, W.I.,(Eds.1995), *Rasch Models, Foundations, Recent Developments and Applications*: Springer-Verlag New York.

Glas, C.A.W. & Verhelst, D.N (1995, pp.69-95). Testing the Rasch Model. In Fischer, H.G. & Molenar, W.I. (Eds. 1995), *Rasch Models, Foundations, Recent Developments and Applications*: Springer-Verlag New York.

Glas, C.A.W. & Verhelst, D.N(1995, pp.325-351). Tests of Fit for polytomous Rasch Models. In Fischer, H.G. & Molenar, W.I., (Eds. 995), *Rasch Models, Foundations, Recent Developments and Applications*: Springer-Verlag New York.

Hambleton, K. R. & Swaminathan, H. (1985, pp.138-140). *Item Response Theory, Principles and Applications*: Kluwer-Nijhoff Publishing.

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

Hemker, T.B. (1996).*Unidimensional IRT models for Polytomous Items, with results for Mokken scale analysis*: NOW.

Masters, N.G. & Wright, D. B. (1997, pp.101-105). The Partial Credit Model. In van der Linden, J. W. & Hambleton, K. R. (Eds. 1997), *Handbook of Modern Item Response Theory*: Springer.

Molenaar, W.I. (1995, pp. 39-51). Estimation of Item Parameters. In In Fischer, H.G. & Molenar, W.I.,(Eds.1995), *Rasch Models, Foundations, Recent Developments and Applications*: Springer-Verlag New York.

Muraki, E. (1997, pp.153-156). A Generalized Partial Credit Model. In van der Linden, J. W. & Hambleton, K. R. (Eds. 1997), *Handbook of Modern Item Response Theory*: Springer.

Verhelst, D.N. & Glas, C.A.W. (1995, pp. 215-237). The One Parameter Logistic Model. In Fischer, H.G. & Molenar, W.I., (Eds.1995), *Rasch Models, Foundations, Recent Developments and Applications*: Springer-Verlag New York.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M.(1995, pp.1-22). *One-Parameter Logistic Model, OPLM*:Cito, National Institute for Educational Measurement, Arnhem.

Chapter 7

Item and Test Information Functions

7.0.0 Introduction

Item response theory provides a method of describing items and tests, selecting test items, assessing precision of measurement and comparing tests. The method applies item information function.

In this chapter item and test information functions and their uses in constructing test will be described with illustrations from the *Project* wherever feasible. As mathematical expression for item and test information functions differ from model to model depending on the number of parameters involved and type of responses involved, only item and test information relevant to OPLM will be dealt in this chapter.

7.1.0 Item Information Function for Dichotomous Model

Hambleton et al., (1991, p.91) present the item information function of dichotomously scored items for one parameter and two parameter logistic models as

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \qquad i = 1,2,\ldots,n. \qquad (7.1.0.1)$$

In equation 7.1.0.1, $I_i(\theta)$ is the information provided by item *i* at θ, $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, $P_i(\theta)$ is the item response function and $Q_i(\theta) = 1 - P_i(\theta)$.

Equation 7.1.0.1 can be verbally defined. According to the equation 7.1.0.1, item information function of dichotomously scored items for one and two parameter logistic models is the ratio of the square of the first derivative of the response probability to the response probability.

Equation 7.1.0.1 is applicable to OPLM when responses are dichotomously scored.

7.2.0 Test Information Function for Dichotomous Model

Hambleton and Swaminathan, (1985, p.104) present the mathematical expression of the test information function for dichotomous model as

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \quad i = 1,2,\ldots,n. \qquad (7.2.0.1)$$

In equation 7.2.0.1, $I(\theta)$ is the test information provided by items *i* through *n* at θ, $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, $P_i(\theta)$ is the item response function and $Q_i(\theta) = 1 - P_i(\theta)$. Equation 7.2.0.1 shows that the information functions are additive (van der Linden (2005, pp. 16-17). According to equation 7.2.0.1 test information function is the sum of the item information functions.

Equation 7.2.0.1 is directly applicable to OPLM.

Equation 7.2.0.1 has the following properties (Hambleton and Swaminathan (1985, p.104):

- *The equation is defined for a set of test items at each point on the ability scale.*
- *The amount of information is influenced by the quality and number of test items.*
- *The steeper the slope the greater the information.*
- *The smaller the item variance the greater the information.*
- *Test information does not depend upon the particular combination of test items. The contribution of each test item is independent of the other items in the test.*
- *The amount of information provided by a set of test items at an ability level is inversely related to the error associated with ability estimates at the ability level*.

7.3.0 Item Information Function for Polytomous Model

Ostini and Nering (2006, p.69) present the mathematical expression of the item information function for polytomous response data as

$$I_i(\theta) = \sum_{g=0}^{m} \frac{[P'_{i_g}(\theta)]^2}{P_{i_g}(\theta)}, \qquad g = 0,1,2\ldots, m. \qquad (7.3.0.1)$$

In equation 7.3.0.1, $I_i(\theta)$ is the information function provided by item *i* at θ, $P'_{i_g}(\theta)$ is the first derivative of the category response probability, $P_{i_g}(\theta)$,of item *i* at g category of *m* categories. Equation 7.3.0.1 defines item information function for a polytomous model as the sum of the information functions of the category response probabilities.

Equation 7.3.0.1 is applicable to OPLM when responses are polytomously scored.

7.4.0 Test Information for Polytomous Model

By using the additive property of the information functions stated in section 7.2.0, test information for polytomous model is the sum of the item information functions across latent trait. Mathematically, test information for polytomous model can be expressed as

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \qquad i = 1,2,\ldots,n. \qquad (7.4.0.1)$$

In equation 7.4.0.1, $I_i(\theta)$ is same as defined in equation 7.3.0.1. Ostini and Nering (2006, p.72, cf. Samejima, 1969) note that the amount of information provided by a polytomous item increases as the number of categories for an item increases.

Equation 7.4.0.1 is applicable to OPLM when responses are scored polytomously.

7.5.0 Test Information Function and Standard Error of Measurement

The test information function is asymptotically equal to the inverse of the variance function of the maximum likelihood estimator of θ (van der Linden (2005, pp. 16-17), meaning that

$$I(\theta) = \frac{1}{Var(\hat{\theta} \mid \theta)}. \qquad (7.5.0.1)$$

Hambleton et al., (1991, pp. 94-95) state that the standard error of $\hat{\theta}$, SE($\hat{\theta}$), is the standard deviation of the asymptotically normal distribution of the maximum likelihood estimate of ability for a given true value of ability θ, meaning that

$$\text{SE}(\hat{\theta}) = \sqrt{Var(\hat{\theta} \mid \theta)} \quad . \qquad (7.5.0.2)$$

From equations 7.5.0.1 and 7.5.0.2, the relationship between test information function and standard error of measurement can be established as

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}, \qquad (7.5.0.3)$$

In equation 7.5.0.3, $SE(\hat{\theta})$ is the standard error of estimation which is equivalent to standard error of measurement in the classical test theory. Equation 7.5.0.3 defines the standard error of measurement as the inverse of the root of the test information function.

Equation 7.5.0.3 is applicable to OPLM. In OPLM, the standard error of measurement is known as the root mean squared error (RMSE) which is expressed as

$$RMSE(t,G) = \left[\int_{\Theta} \int_{T} (t - \tau(\theta))^2 \, dH(t) \mid \theta) dG(\theta) \right]^{\frac{1}{2}}, \qquad (7.5.0.4)$$

In equation 7.5.0.4 (Verhelst et al., 1995, p. 99), Θ represents the domain of θ and $\tau(\theta)$ is the function of θ, *t* is an estimator of τ with domain *T*, $H(t \mid \theta)$ is the conditional distribution of *t* given by θ and $G(\theta)$ is the distribution of θ.

Equation 7.5.0.4 is used to define accuracy of measurement in OPLM as

$$MAcc = \frac{Var(\tau(\theta)}{Var(\tau(\theta) + MSE(t,G)}, \qquad (7.5.0.5)$$

where $MSE = RMSE^2$.

From equation 7.5.0.3, the factors influencing the standard errors are identifiable by using the features of the test information function. Hambleton et al., (1991, pp. 94-95) present the following factors on which the magnitude of standard error depends:

- *The number of test items*: The longer the length of a test, the smaller the standard error.

- *The quality of the test items*: Smaller standard errors are associated with highly discriminating items for which the correct answers cannot be obtained by guessing.

- *The match between item difficulty and examinee ability*: Smaller standard errors are associated with tests composed of items with difficulty parameters approximately equal to the ability parameter of the examinee.

7.6.0 Summary

Item response theory provides a method of describing items and tests, selecting test items, assessing precision of measurement and comparing test. The method applies item information function.

Item information function of dichotomously scored items for one and two parameter logistic models is the ratio of the square of the first derivative of the response probability to the response probability. Test information function is the sum of the item information functions.

For a polytomous model, item information function is the sum of the information functions of the category response probabilities. Test information function is the sum of the item information functions.

The standard error of measurement is the inverse of the root of the test information function. Standard error of measurement depends on (a) the number of test items, (b) the quality of the test items and (c) the match between item difficulty and examinee ability.

7.7.0 References

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

Van der Linden, J. W. (2005). *Linear Models for Optimal Test Design*: Springer.

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

Ostini. R. & Nering, L. M. (2006). *Polytomous Item Response Theory Models*: Sage Publications.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995, pp.1-22). *One-Parameter Logistic Model, OPLM*:Cito, National Institute for Educational Measurement, Arnhem.

Chapter 8

Item Calibration

8.0.0 Introduction

Item calibration is a method of defining item parameters for an item or for a group of items. When CTT is used to calibrate items, the item parameters are (a) p-values (alias: item easiness), (b) item discrimination index, and (c) point bi-serials. The item parameters obtained from CTT are population dependent. The values of item parameters will change when examinees change. This is a strong weakness of the CTT.

When IRT is used to calibrate item parameters, the number of item parameters available to a test designer will depend on the type of IRT model used for calibration. For instance, One Parameter Logistic Model has an item difficulty but not item discrimination (assumed to be uniform with value 1) and Generalized Partial Credit Model has both item difficulty and item discrimination. The interpretation of item parameters will also differ from model to model in IRT. The values of the item parameters obtained from IRT are population invariant. This gives IRT a superior strength over CTT.

In this chapter, item calibration for test items used in UIBTERV will be described.

8.1.0 Type of IRT Model used for Item Calibration

OPLM is used as the IRT model for calibrating the test items used in UIBTERV. The OPLM model is presented in chapter 6 as equation 6.1.0.1. With OPLM, the kinds of item parameters generated are (a) item difficulty and (b) item discrimination. Since response data used in UIBTERV is dichotomous response data, the OPLM takes the form of equation 6.5.2.1 of chapter 6. In equation 6.5.2.1, item discrimination parameter is represented by *a* and item difficulty parameter is represented by *b*.

The value of item discrimination parameter shows how an item is able to differentiate high ability examinees and low ability examinees. An item with high item discrimination parameter is more discriminative than an item with low item discrimination parameter. In other words, item discrimination parameter measures the worth of an item in terms of its usefulness in separating high ability examinees and low ability examinees. It is in the light of this property of the item discrimination parameter that a psychometrician gets delighted by an item with high item discrimination value and concerned by an item with low item discrimination value.

The value of item difficulty parameter shows the difficulty of an item, though. The item difficulty parameter is a location on the ability continuum where an examinee has fifty percent chance of getting the item correct and fifty percent chance of getting the item incorrect. This point is illustrated in figure 8.1.0.1.

The item difficulty parameter is useful in interpreting how the examinees with different abilities will respond to an item. The examinees whose abilities fall before the ability that correspond to the item difficulty parameter on the ability continuum are expected to score the item incorrectly. The examinees whose abilities exactly fall on the location of the ability that corresponds to the item difficulty parameter on the ability continuum have fifty percent chance of scoring the item correctly and fifty percent chance of scoring the item incorrectly. The examinees whose abilities fall after the ability that corresponds to the item difficulty parameter on the ability continuum are expected to score the item correctly.

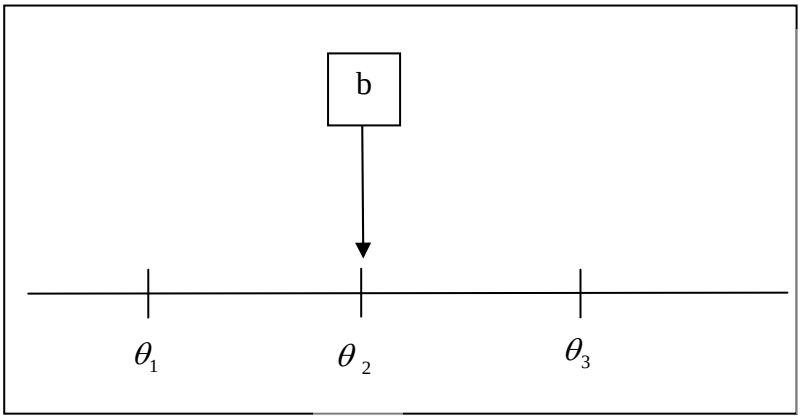

Figure 8.1.0.1: Shows item difficulty parameter, *b*, as the location parameter on the ability continuum. Examinees with ability θ_1 will fail the item and examinees with ability θ_2 will have 50 percent chance of responding the item correctly and examinees with ability θ_3 will find the item very easy.

8.2.0 Steps of Item Calibration

After selecting appropriate model from the available IRT models, the goodness of the fit of the model to the response data has to be established first. If the goodness of fit of the model to the response data is poor, item parameters generated by using the model will be faulty and misleading. The OPLM has a number of fit statistics both at item level and test level and they are presented in chapter 6. Once the goodness of fit of the chosen model to the response data is established, item parameters can be generated by the model.

8.2.1 Fitting OPLM to UIBTERV Data

The data file used from UIBTERV is *EngEntNw.Dat*. The data file consists of 358 items written in 13 test booklets. The details of the data are presented in section 1.2.1 of chapter 1. Each test booklet has 70 items with 35 items measuring reading compression construct and other 35 items measuring vocabulary construct. The number of respondents is 1280. The test booklets are linked to one another by common items called anchor items. These anchor items ensure uniform scale across the test booklets.

To perform the goodness of fit analysis, the data will have to be read into the OPLM. The steps involved in reading data into the OPLM are as follows. Read the data *EngEntNw.Dat* into the OPLM by using **file** button. Specify the **JobNm** and the number of **items.** Define item labels which distinguish the reading items and vocabulary items. A unique file for item labels has to be predefined in accordance with the data structure. In UIBTERV, item labels are defined in *Entry i.txt* file. The item label file is read into the OPLM by using **Item label** command: Click **File**, select **Read Text**, select **Item Array** and click **Item Labels.** After clicking **Item Label**, file containing item labels will be asked. Enter *Entry i.txt* and item labels are read into the OPLM. Next the booklet identity numbers have to be read into the OPLM. Based on the position of the column occupied by the booklet identity numbers, column number and length of the booklet identity number have to be specified in **Column** and **Length.** In *EngEntNw.Dat*, booklet ID begins from column 51 and has 2 digits, i.e. **Column** is 51 and **Length** is 2. The beginning of the column with response data will have to be specified in **First Kolumn** and the size of the response in **Length of Response**. In *EngEnt.Dat*, the response data begins from column 61 and response has one digit score, i.e. **First Kolumn** is 61 and **Length of Response** is 1. The **General** part of the OPLM screen file is ready for analysis. The type of analysis will have to be specified by selecting options under **Analysis.** For details on the options, a reference may be made to Verhelst, et al., (1995).

Before analysis can be performed, **2Design** has to be completed. The number of booklets has to be identified in **Booklet.** In case of the UIBTERV, the number of booklets is 13 originally, but for analyses these 13 booklets are transformed into 28 booklets. Items in each booklet have to be read into **ItemId's in Booklet.** A file containing the items in each booklet has to be written. This file has to be

pre-designed. In case of UIBTERV, the file *Entry.b.txt* contains the items in each booklet. *Entry.b.txt* is read into **ItemId's in Booklet** as follows: Click on **File**, click **Design** and click **Booklets**. A pop up will ask for the file containing items in the booklets. In case of UIBTERV, *Entry.b.txt* is read into **ItemId's in Booklet**. The screen file for UIBTERV up to this point is *Eng_AOL.SCR*.

8.2.2 Analysis 1

The OPLM is ready for the goodness of fit analysis. To study if two constructs, viz., reading comprehension and vocabulary can be fitted into the model, an analysis is made involving both the constructs. It is found that the fit was very poor. This is not a surprise because OPLM has strong assumption of unidimensionality of latent trait. As a result, the reading comprehension items have to be separated from the vocabulary items.

To separate reading comprehension items and the vocabulary items, a file has to be created which defines the status of items in calibration. Items can be allotted two statuses, viz., On and Off. The On and OFF are defined by binary digits 1 and 0 respectively. The file containing item labels can be used for making new files that define the status of items in calibration. In case of UIBTERV, two files are made by using *Entry i.txt*. File *EntryRe.oo.txt* contains reading items with ON status and file *EntryW.oo* contains items with vocabulary items with ON status. The Screen file *Eng_AOL.SCR* is used to make two screen files. One screen file contains only reading items with ON status and the other screen file contains only vocabulary items with On status. To make the screen file with reading items, click on **File**, click **Item Array** and click **Calibrate ON/OFF**. A pop up will ask for the item status file. In case of UIBTERV, *EntryRe.oo.txt* is read into the OPLM. The new **JobNm** of the new screen is defined as *ER_0.SCR*. Similar steps is applied for making screen file *NER_0.SCR* by using *EntryW.oo.txt* file.

8.2.3 Analysis 2 :Goodness of Fit Test

In analysis 2, English Reading Comprehension Test is analyzed. The 1st step in the analysis is to perform Rasch analysis, where OPLM functions as the Rasch Model. The screen file used is NER__0.SCR. The main purpose of the Rasch analysis is to assess the general fit of the model to the data. Since the Rasch model is a strong model, many items do not fit the model. As a result, OPLM is used.

In the 2nd step, screen file *NER__1.SCR* is made. The item discrimination parameter is imputed by using running **Opcat** module. The mean of the geometric progression is set at 2. Series of analyses are made till a good fit is obtained. Table 8.2.3.1 presents the analyses.

Analysis		Item #	Misfit causes	Screen File	# Items analyzed	# of Items out of range	Total Items
Round	Type						
1	Rasch	11	Nonmonotonicity	NER_0.SCR	144	19	163
		52	Nonmonotonicity				
2	OPLM	55	Flat for high ability group	NER_1.SCR	141	22	163
		215	Guessing & nonmonotonicity				
3	OPLM	146	Nonmonotonicity and flat for high ability group	NER_2.SCR	139	22	161
		179	Nonmonotonicity				
4	OPLM	42	Flat	NER_3.SCR	137	22	159

Analysis		Item #	Misfit causes	Screen File	# Items analyzed	# of Items out of range	Total Items
Round	Type						
		132	Nonmonotonicity				
5	OPLM	Duplicate of NER_3.SCR		NER_4.SCR	137	22	159
6	OPLM	197	Large deviance	NER_5.SCR	135	22	157
7	OPLM		No misfitting item	NER_6.SCR	134	22	156

Table 8.2.3.1: Shows the results of the analyses.

The misfitting items are put off from the calibration. In round 7, the goodness of fit is established and the scale for further analyses is fixed. The fit statistics obtained from different rounds are shown in figure 8.2.3.2.

Once the scale is fixed, the item parameters are fixed. The calibration of the items is complete. However, as can be seen in table 8.2.3.1, there are 22 items out of range. The 22 items did not have S-statistic because of very high p- values. To involve the 22 items in the calibration, Differential Item Functioning (DIF) has to be studied. Table 8.2.3.3 shows the 22 items. The next section of analyses will perform DIF analysis.

8.3.0 Analysis 3 : Differential Item Functioning

Differential item functioning (DIF) is said to occur whenever examinees from two different population groups that have the same amount of the underlying trait measured by the test perform unequally on an item (Bolt, 2002; Clause & Mazor, 1998; Montesino & Lopen-Pina, 2002). To facilitate DIF study, the examinees are usually categorized into a reference group and a focal group. These groups are DIF populations. Performance of focal group is compared with the performance of the reference group. DIF analysis involves testing null hypotheses in the light of the performances of the DIF populations as follows:

H0: The item functions equally for the reference and focal groups (no DIF).
HA: The item functions unequally for the reference and focal groups (DIF).

Different methods of testing DIF hypotheses are available subject to how DIF populations are defined. In UIBTERV, DIF analyses are done by using the OPLM. The OPLM has a module called **OPDRAW** which displays three kinds of plots, viz., (a) item information function, (b) category response curves and the regression of the item information on the latent variable and (c) simultaneous plot of observed proportions and expected probabilities. To perform DIF analyses, option (c) is used.

In UIBTERV, DIF analyses are done at two aggregation levels, viz., (a) gender and (b) school.

For example, figure 8.3.0.1 shows a DIF item number 88 for two DIF populations: Male and Female. When the curve of the observed proportion falls out of the 95% confidence envelope, as is the case with item number 88, the item is detected as functioning differentially across different populations. If the curve of the observed proportion is within the 95% confidence envelope for both male and female populations, the item does not function differentially across populations, as is the case with the item shown in figure 8.3.0.2. The middle blue lines (subject to color availability) in the plots are the curves of the expected proportions.

Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 10/ 16/ 11 13 16 13 16 15 10 10 7 7 R1c* = 993.137; df = 708; p = .0000	Round 1
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 1/ 6/ 4 21 12 16 13 10 17 15 11 15 R1c* = 837.133; df = 737; p = 0.0056	Round 2
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 0/ 2/ 8 12 15 8 16 13 20 18 13 12 R1c* = 816.160; df = 711; p = .0034	Round 3
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 0/ 2/ 8 12 15 8 16 13 20 18 13 12 R1c* = 816.160; df = 711; p = .0034	Round 4
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 0/ 2/ 8 12 15 8 16 13 20 18 13 12 R1c* = 816.160; df = 711; p = .0034	Round 5
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 0/ 2/ 4 15 14 11 22 10 16 21 7 13 R1c* = 785.611; df = 698; p = .0110	Round 6
Distribution of p-values for S-tests. 0.--/---/---.1-----.2-----.3-----.4-----.5-----.6-----.7-----.8-----.9-----1. 0/ 2/ 7 14 15 14 15 11 15 20 7 14 R1c* = 794.044; df = 717; p = .0230	Round 7

Figure 8.2.3.2: Fit statistics obtained from different rounds of analyses.

Figure 8.3.0.1 shows a DIF item number 88 for two DIF populations: Male and Female.

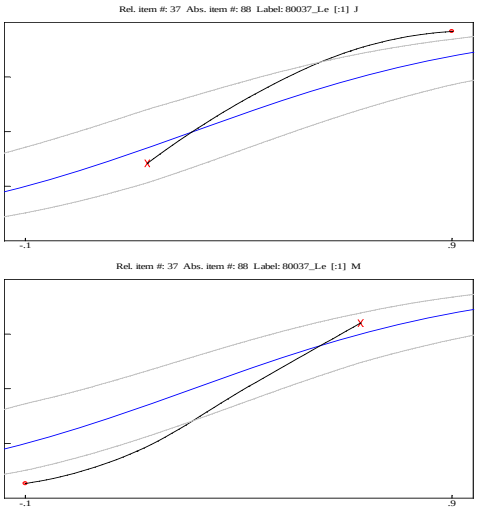

Figure 8.3.0.1: Shows DIF item

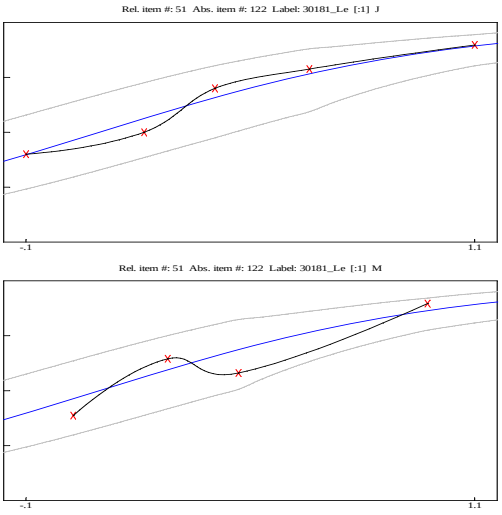

Figure 8.3.0.2: Shows an item without DIF

8.3.1 Gender DIF Analysis

The two populations defined for DIF are female and male. To calibrate items separately for these two DIF populations, new booklets have to be assigned to them. This is done by using **Crosstab** function from the *SPSS 12*. Table 8.3.1.1 shows the number of males and females in different booklets.

By using table 8.3.1.1, the new booklets for unknown population, male population and female population are defined as in table 8.3.1.2.

Booklet	Unknown	Gender		Total	Booklet	Unknown	Gender		Total
		J	M				J	M	
1	1	55	47	103	15	0	7	7	14
2	0	40	39	79	16	1	2	7	10
3	2	42	48	92	17	0	8	7	15
4	4	84	63	151	18	0	4	4	8
5	1	83	72	156	19	0	1	1	2
6	0	33	32	65	20	0	6	5	11
7	1	79	71	151	21	0	5	7	12
8	1	54	51	106	22	0	8	10	18
9	2	54	57	113	23	0	0	1	1
10	0	37	33	70	24	0	5	6	11
11	4	30	16	50	25	0	1	0	1
12	1	26	18	45	26	0	13	3	16
13	2	17	28	47	27	1	2	6	9
14	0	7	6	13	28	0	5	5	10

Table 8.3.1.1: Shows results of the cross tabulation of gender and booklets.

Original Booklet	Mapped to New			Original Booklet	Mapped to New
	Booklet for Unknowns	Booklet for Male	Booklet for Female		Booklet for Unknown
1	99	1	14	15	99
2	99	2	15	16	99
3	99	3	16	17	99
4	99	4	17	18	99
5	99	5	18	19	99
6	99	6	19	20	99
7	99	7	20	21	99
8	99	8	21	22	99
9	99	9	22	23	99
10	99	10	23	24	99
11	99	11	24	25	99
12	99	12	25	26	99
13	99	13	26	27	99
14	99			28	99

Table 8.3.1.2: Shows booklets assigned to male and female.

The minimum number of examinees in each booklet is fixed at 14. The new booklets have to be defined in the original data with appropriate transformation. After the new booklets are defined in the data file, they are read into the OPLM by using **2Design** command: Click **2Design**, fill **Booklets** with the number of booklets, fill **Populations** with the number of populations, fill **DIF Values** with number of DIF populations and click **Copy Booklets** and fill the boxes with appropriate information. In case of *English Reading Comprehension Test*, number of booklets is 26, number of populations is 2 (defined as Male and Female) and DIF value is 2 (defined as Male and Female) and booklets 1-13 are transformed to booklets 14-26 for female population. Since the original positions of the booklets and the response data usually are changed, i.e. their original columns are shifted to either right or left, the booklet positions and the response data positions will have to be carefully defined according to their new positions in the data. The screen file for doing DIF analyses is ready. In case of UIBTERV-*English Reading and Comprehension Test*, the first DIF screen file is NER*DG.SCR*.

The result of the DIF analysis for gender population is displayed in table 8.3.1.3.

Analysis		Item #	Misfit causes	Screen File	# Items analyzed	# of Items out of range	Total Items
Round	Type						
1	OPLM	78	RP of male less than RP of female	NERDG.SCR	156	Nil	156
		88	RP of female less than RP of male				
		149	RP of female less than RP of male				
2	OPLM	353	RP of male more than RP of female	NERDG_1.SCR	153	Nil	153
3	OPLM	Nil	No DIF item	NERDG_2.SCR	153	Nil	153

Table 8.3.1.3: Shows 3 DIF items.

8.3.2 School Level DIF Analyses

Similar steps are applied to perform DIF analyses for school levels. The school level has 7 possible DIF populations, viz., Unknown, GT, KB, BB, BB+, HAVO and VWO. However, only 5 DIF populations are made for DIF analyses. The 5 DIF populations and their DIF values are shown in table 8.3.2.1.

Populations	DIF Values
Unknown	Unknown
BB/BB+	BB/BB+
KB	KB/GL
GL	KB/GL
HA/VW	HA/VW

Table 8.3.2.1: Shows the DIF populations and DIF values for school level populations.

Analysis		Item #	Misfit causes	Screen File	# Items analyzed	# of Items out of range	Total Items
Round	Type						
1	OPLM	280	RP of HA/VW greater than RP of Unknown	NERDS.SCR	156	Nil	156
		345	RP of Unknown greater than RP of HA/VW and RP of HA/VW greater than RP of KB/GL				
2	OPLM	22	RP of BB/BB+ less than RP of other populations	NERDS1.SCR	152	Nil	152
3	OPLM	Nil	No DIF item	NERDS3.SCR	151	Nil	151

Table 8.3.2.2: The results of the analyses of school level DIF populations.

The item calibration for English Reading Comprehension Test is completed. The final results of the calibration are as shown in table 8.3.2.3.

Nr	label	A	B	SE(B)	S*	DF	Probability	Percent Correct	Remarks
8	0060_Le	2	0.439	0.08	10.012	6	0.124	0.584	
9	0016_Le	2	-0.022	0.086	5.626	5	0.344	0.753	
10	0019_Le	3	0.121	0.067	0.61	4	0.962	0.749	
11	0012_Le	1	1.399	0.145	2.043	7	0.957	0.33	
12	0022_Le	2	-1.066	0.155	0.014	1	0.906	0.955	
21	0190_Le	2	-0.294	0.095	2.076	4	0.722	0.831	
22	0209_Le	1	-0.857	0.163	8.295	6	0.217	0.809	School level DIF
23	0014_Le	2	0.43	0.08	3.403	6	0.757	0.588	
24	0018_Le	2	-0.022	0.086	1.527	5	0.91	0.753	
25	0019_Le	2	0.121	0.083	3.262	5	0.66	0.704	
31	40031_Le	2	-0.945	0.141	0.413	1	0.521	0.944	

Nr	label	A	B	SE(B)	S*	DF	Probability	Percent Correct	Remarks
32	40049_Le	2	-0.455	0.103	1.733	4	0.785	0.869	
33	40058_Le	2	-0.785	0.125	2.933	2	0.231	0.925	
34	40062_Le	2	-0.91	0.137	4.179	2	0.124	0.94	
35	40080_Le	3	0.013	0.07	9.085	4	0.059	0.794	
41	40001_Le	3	-0.431	0.103	2.373	1	0.123	0.922	
43	40003_Le	2	-1.093	0.178	0	0	99.999	0.956	
44	40006_Le	2	-1.403	0.232	0	0	99.999	0.976	
45	40007_Le	2	0.34	0.093	5.771	4	0.217	0.62	
51	80004_Le	3	-0.053	0.082	0.682	3	0.878	0.815	
52	80006_Le	1	0.345	0.159	6.94	5	0.225	0.556	
53	80008_Le	3	0.169	0.076	1.07	3	0.784	0.722	
54	80005_Le	2	0.046	0.097	2.511	4	0.643	0.727	
60	40009_Le	2	-0.013	0.098	5.635	4	0.228	0.746	
61	40035_Le	2	-0.316	0.109	4.255	3	0.235	0.834	
62	40050_Le	2	-0.105	0.101	5.757	4	0.218	0.776	
63	40066_Le	2	0.69	0.093	5.783	4	0.216	0.483	
64	50062_Le	2	-0.278	0.108	9.309	3	0.025	0.824	
75	40011_Le	2	0.015	0.098	1.294	3	0.731	0.713	
76	40015_Le	2	-0.703	0.135	0.88	1	0.348	0.901	
77	40023_Le	3	-0.539	0.111	1.008	1	0.315	0.928	
78	40025_Le	2	-1.265	0.183	0	0	99.999	0.974	Gender DIF
79	40028_Le	2	-1.407	0.232	0	0	99.999	0.972	
85	80026_Le	2	0.092	0.097	2.068	3	0.558	0.685	
86	80028_Le	2	0.307	0.094	1.81	4	0.771	0.602	
87	80029_Le	2	-0.281	0.108	5.956	3	0.114	0.807	Gender DIF
88	80037_Le	3	0.279	0.074	7.601	3	0.055	0.635	
89	80027_Le	2	0.673	0.094	2.832	4	0.586	0.453	
95	40029_Le	3	-0.43	0.101	0.846	1	0.358	0.906	
96	40030_Le	3	-0.641	0.122	0	0	99.999	0.945	
97	50056_Le	3	0.108	0.077	2.433	3	0.487	0.718	
98	50075_Le	2	0.645	0.094	4.747	4	0.314	0.464	
99	40076_Le	2	-0.891	0.153	0.559	1	0.455	0.928	
110	40018_Le	2	-0.591	0.088	1.623	5	0.898	0.896	
111	40010_Le	2	0.459	0.063	5.976	7	0.543	0.572	
112	40021_Le	2	-0.131	0.064	9.888	6	0.129	0.815	
113	40032_Le	1	-0.278	0.113	2.561	7	0.922	0.706	
114	40036_Le	2	-0.437	0.081	2.496	5	0.777	0.866	
120	30078_Le	3	-0.27	0.065	0.834	4	0.934	0.891	
121	30183_Le	3	0.128	0.054	4.165	5	0.526	0.752	
122	30181_Le	2	0.096	0.066	3.928	7	0.788	0.713	
123	0023_Le	3	0.549	0.051	5.714	6	0.456	0.542	
124	0021_Le	2	0.203	0.065	7.086	7	0.42	0.674	
130	0046_Le	1	-1.089	0.137	7.249	7	0.403	0.84	
131	0048_Le	2	0.576	0.063	7.103	7	0.418	0.523	
133	50038_Le	3	0.01	0.056	3.516	5	0.621	0.801	
134	50077_Le	2	0.609	0.063	4.213	7	0.755	0.509	

Nr	label	A	B	SE(B)	S*	DF	Probability	Percent Correct	Remarks
145	40004_Le	2	-0.641	0.113	3.393	3	0.335	0.93	
147	40017_Le	2	-1.565	0.252	0	0	99.999	0.988	
148	40033_Le	2	-0.692	0.118	5.211	2	0.074	0.936	
149	40056_Le	1	-0.646	0.143	4.285	7	0.746	0.802	Gender DIF
157	30162_Le	2	-0.154	0.083	7.344	5	0.196	0.842	
158	30220_Le	1	-1.099	0.164	5.498	6	0.482	0.863	
159	30128_Le	1	-0.902	0.154	4.421	6	0.62	0.839	
160	30133_Le	4	0.094	0.054	5.219	3	0.156	0.872	
161	30150_Le	2	0.551	0.067	3.849	7	0.797	0.605	
169	50001_Le	1	0.306	0.12	16.742	7	0.019	0.62	
170	50078_Le	2	0.116	0.074	5.286	6	0.508	0.766	
171	50079_Le	2	0.087	0.075	11.561	6	0.073	0.775	
172	50052_Le	2	-0.205	0.085	6.009	5	0.305	0.854	
173	50071_Le	3	0.295	0.055	10.452	5	0.063	0.757	School level DIF
180	40054_Le	2	-0.202	0.09	3.232	4	0.52	0.871	
181	40057_Le	3	0.197	0.062	3.397	4	0.494	0.826	
182	40072_Le	2	0.357	0.072	7.238	6	0.299	0.715	
183	50024_Le	2	-0.147	0.087	8.252	5	0.143	0.859	
189	30250_Le	2	-0.083	0.084	7.674	5	0.175	0.844	
190	30249_Le	2	0.118	0.077	10.244	5	0.069	0.791	
191	30330_Le	3	-0.011	0.07	0.339	3	0.952	0.888	
192	30394_Le	2	-0.023	0.082	7.079	5	0.215	0.829	
193	30270_Le	2	0.439	0.07	7.688	6	0.262	0.685	
196	50031_Le	1	-0.454	0.138	2.598	7	0.92	0.785	
198	50068_Le	3	0.001	0.069	2.608	3	0.456	0.885	
199	50063_Le	3	0.323	0.058	5.329	5	0.377	0.779	
200	50047_Le	2	-0.824	0.137	0.481	1	0.488	0.956	
212	40053_Le	2	0.34	0.072	3.684	6	0.719	0.735	
213	40064_Le	1	0.328	0.121	6.298	7	0.505	0.647	
214	40077_Le	1	-1.202	0.176	6.827	5	0.234	0.888	
216	50030_Le	2	0.473	0.07	3.71	6	0.716	0.688	
222	30246_Le	4	0.179	0.058	4.675	3	0.197	0.885	
223	30245_Le	2	-0.112	0.087	3.373	5	0.643	0.862	
224	30226_Le	2	0.357	0.072	5.571	6	0.473	0.729	
225	30227_Le	2	0.102	0.079	3.32	6	0.768	0.809	
226	30247_Le	3	0.206	0.063	6.475	4	0.166	0.841	
232	50057_Le	3	1.048	0.055	3.203	5	0.669	0.447	
233	50040_Le	3	0.079	0.068	3.237	3	0.357	0.879	
234	50041_Le	2	-0.615	0.121	0.068	2	0.967	0.941	
235	50042_Le	2	0.174	0.077	8.551	6	0.2	0.788	
236	50043_Le	3	0.09	0.067	0.261	3	0.967	0.876	School level DIF
250	40045_Le	2	0.865	0.093	1.337	4	0.855	0.596	
251	40055_Le	2	0.28	0.108	6.653	3	0.084	0.803	
252	40068_Le	3	0.185	0.098	1.159	1	0.282	0.893	
253	40070_Le	3	-0.279	0.15	0	0	99.999	0.966	
254	40071_Le	1	-0.079	0.184	0.913	3	0.822	0.758	

Nr	label	A	B	SE(B)	S*	DF	Probability	Percent Correct	Remarks
262	30280_Le	2	0.357	0.105	1.725	3	0.631	0.781	
263	30196_Le	2	0.755	0.094	1.526	3	0.676	0.64	
264	30125_Le	1	-0.68	0.216	1.745	3	0.627	0.848	
265	30098_Le	2	0.24	0.11	1.692	2	0.429	0.815	
266	30053_Le	2	0.608	0.097	3.561	3	0.313	0.697	
271	40073_Le	1	-0.636	0.213	0.223	3	0.974	0.843	
272	50070_Le	3	0.081	0.106	0.51	1	0.475	0.916	
273	50039_Le	2	-0.431	0.164	0.063	1	0.802	0.938	
274	40051_Le	1	-0.773	0.222	3.318	3	0.345	0.86	
275	50044_Le	2	0.106	0.117	0.462	2	0.794	0.848	
276	40008_Le	1	-0.494	0.24	0.412	2	0.814	0.832	
277	40034_Le	3	0.306	0.106	0	0	99.999	0.878	
278	40038_Le	2	-0.188	0.167	0	0	99.999	0.916	
279	40044_Le	1	0.881	0.163	3.301	4	0.509	0.582	
280	30325_Le	1	-0.611	0.249	0.896	2	0.639	0.847	School level DIF
281	30342_Le	1	-0.14	0.218	4.052	2	0.132	0.779	
282	30462_Le	3	0.48	0.096	0.009	1	0.924	0.824	
283	30172_Le	2	0.209	0.132	0.002	1	0.968	0.84	
284	30130_Le	2	0.38	0.122	1.385	2	0.5	0.794	
291	40059_Le	2	-0.369	0.19	0	0	99.999	0.939	
292	40075_Le	2	-0.369	0.19	0	0	99.999	0.939	
293	50080_Le	3	0.306	0.106	0	0	99.999	0.878	
294	50059_Le	2	-0.004	0.148	0.155	1	0.694	0.885	
305	40037_Le	2	0.23	0.15	0.092	1	0.761	0.823	
306	40039_Le	2	1.508	0.13	0.223	1	0.637	0.344	
307	40042_Le	2	0.342	0.143	1.042	1	0.307	0.792	
308	50066_Le	3	0.046	0.147	0	0	99.999	0.927	
314	17035_Le	4	0.653	0.092	0	0	99.999	0.76	
315	21042_Le	2	0.055	0.164	0	0	99.999	0.865	
316	21055_Le	1	0.004	0.246	0.198	1	0.656	0.75	
317	30081_Le	2	0.712	0.129	0.089	1	0.766	0.667	
318	30140_Le	3	0.78	0.101	0.158	1	0.691	0.677	
323	50067_Le	3	0.865	0.1	1.311	1	0.252	0.635	
324	50037_Le	2	0.712	0.129	0.089	1	0.766	0.667	
325	40074_Le	2	0.684	0.13	0.01	1	0.921	0.677	
326	50050_Le	3	0.269	0.123	0	0	99.999	0.875	
327	50054_Le	2	0.508	0.135	0.528	1	0.467	0.74	
330	40040_Le	2	-0.343	0.221	0	0	99.999	0.944	
331	40043_Le	2	-0.565	0.264	0	0	99.999	0.963	
332	40047_Le	2	0.157	0.159	2.239	1	0.135	0.87	
333	40060_Le	1	0.596	0.219	2.166	2	0.339	0.648	
334	40065_Le	2	0.157	0.159	2.239	1	0.135	0.87	
342	01115_Le	3	0.852	0.098	0.597	1	0.44	0.713	
343	04038_Le	2	0.008	0.174	0	0	99.999	0.898	
344	30060_Le	2	0.46	0.137	0.095	1	0.758	0.796	
345	30142_Le	2	0.718	0.126	0.981	2	0.612	0.713	

Nr	label	A	B	SE(B)	S*	DF	Probability	Percent Correct	Remarks
346	30168_Le	1	-0.735	0.294	0.123	1	0.726	0.87	
353	50026_Le	2	0.582	0.131	0.342	1	0.559	0.759	Gender DIF
354	50055_Le	2	0.492	0.136	0.921	1	0.337	0.787	School level DIF
355	50045_Le	2	1.068	0.118	0.336	2	0.845	0.574	
356	50035_Le	2	-0.256	0.207	0	0	99.999	0.935	
357	50034_Le	3	0.388	0.122	0	0	99.999	0.88	
358	=E279_Le	1	0.524	0.307	0	0	99.999	0.62	

Table 8.3.2.3: Shows the results of item calibration English Reading Comprehension.

8.4.0 Summary

Item parameters obtained by using CTT are population dependent where as item parameters obtained by using IRT are population invariant.

The item discrimination shows how an item is able to discriminate low ability examinees and high ability examinees. An item difficulty parameter is the location on the ability continuum where an examinee has 50 percent chance of getting the item correct and 50 percent chance of getting the item incorrect.

A poor goodness of fit of the model to the response data will give misleading item parameters.

Differential item functioning is said to occur whenever examinees from two different population groups that have the same amount of the underlying trait measured by the test perform unequally on an item. DIF analyses involve testing null hypotheses.

8.5.0 References

Bolt, M.D. (2002). A Monte Carlo Camparison of Parametric and Nonparametric polytomous DIF Detection Methods. In *Applied Measurement Education, 15 (2), 113-141*: Lawrence Erlbaum Associates, Inc.

Clauser, E.M. & Mazor, M.K. (1998). An NCME Instructional Module on Using Statistical Procedures to Identify Differentially Functioning Test Items. In *Educational Measurement, ITEMS*: National Council on Measurement in Education, Princeton, NJ.

Montesinos, H.D.M. & lopen-Pina, A.J. (2002). Two Stage Equating in Differential Item Functioning Detection under the graded Response Model with the Raju Area Measures and the Lord Statistic. In *Educational & Psychological Measurement, Vol.62,No.1, February 2002, 32-44*: Sage Publications.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995, pp.1-22). *One-Parameter Logistic Model, OPLM*:Cito, National Institute for Educational Measurement, Arnhem.

Chapter 9

Generating Item Information and Global Norms

9.0.0 Introduction

The purposes of norms, the definition of normative scores and the methods involved in making norm reference table are presented in chapter 7. The OPLM also generates expected item information for each population as well as item p-values along with norm table. The expected item information and item p-values are used for selecting items for making tests.

In this chapter, procedures for producing global norm and expected item information will be discussed.

9.1.0 Global Norm

The term global norm as used in the chapter will denote the norm reference table built by using all response data available for English Reading Comprehension from UIBTERV. The global norm reference table uses percentile ranks as normative sores to indicate the percentage of examinees in a norm group who have trait value less than or equal to that particular trait value.

To build global norm, **OPLAT** module of the OPLM is used. For detailed description of **OPLAT** module, a reader may be interested to read Verhelst et al., (1995, p.89).

9.1.1 Population Parameters

To generate global norm table of reference, population parameters are required. In UIBTERV English Reading Comprehension, the types of populations identified are (a) ALL, (b) Gender and (c) School Level. The population parameters for these populations are generated by running **OPMML** module of the OPLM.

9.1.2 Estimation of Population Parameters

The estimation of population parameters have to be done several times based on different conditions applied to the data. Usually the conditions are made by including and excluding misfitting items and DIF items from the data. In UIBTERV English Reading Comprehension, the population parameters are estimated under the conditions shown in table 9.1.2.1.

Sl.N0.	Condition
1	Without misfitting and DIF items across all populations
2	Without misfitting items, school level DIF items and gender DIF items across gender populations
3	Without misfitting items, gender DIF items and school level DIF items across school level populations
4	Without misfitting items but with gender DIF items and school level DIF items across all populations
5	Without misfitting items but with gender DIF items and school level DIF items across gender populations
6	Without misfitting items but with gender DIF items and school level DIF items across school level populations
7	With less severe misfitting items and with all DIF items across all populations
8	With less severe misfitting items and with all DIF items across gender populations
9	With less severe misfitting items and with DIF items across school level populations

Table 9.1.2.1: Shows the conditions under which population parameters are estimated.

Based on the conditions shown in table 9.1.2.1, the population parameters generated by using the **OPMML** module are shown in table 9.1.2.2.

C	P	MML Estimates of Population Parameters		Files	
		Mean(S.E)	S.D(S.E)	SCREEN	PAR
1	ALL	0.824(0.016)	0.523(0.014)	NERAA_1.SCR	NERAA.PAR
2	Gender				
	Male	0.852(0.024)	0.544(0.021	NERAA_3.SCR	NERAA.PAR
	Female	0.807(0.024)	0.493(0.020)		
3	School Level				
	Unknown	0.878(0.025)	0.520(0.022)	NERAA_5.SCR	NERAA.PAR
	BB+/BB	0.536(0.034)	0.459(0.030)		
	KB	0.727(0.044)	0.509(0.039)		
	GL	0.795(0.034)	0.430(0.031)		
	HA/VW	1.154(0.044)	0.484(0.038)		
4	ALL	0.836(0.016)	0.522(0.014)	NER_7.SCR	NER_6.PAR
5	Gender				
	Male	0.869(0.024)	0.547(0.021)	NERDDG.SCR	NER_6.PAR
	Female	0.813(0.023)	0.490(0.020)		
6	School Level				
	Unknown	0.892(0.025)	0.515(0.022)	NERDDS.SCR	NER_6.PAR
	BB+/BB	0.548(0.034)	0.459(0.030)		
	KB	0.730(0.042)	0.495(0.038)		
	GL	0.799(0.044)	0.430(0.030)		
	HA/VW	1.176(0.044)	0.489(0.038)		
7	ALL	0.836(0.016)	0.523(0.014)	NERLSDDM.SCR	NERLSDD.PAR
8	Gender				
	Male	0.869(0.024)	0.547(0.021)	NERLDDG.SCR	NERLSDD.PAR
	Female	0.812(0.023)	0.490(0.020)		
9	School Level				
	Unknown	0.892(0.025)	0.515(0.022)	NERLDDXS.SCR	NERLSDD.PAR
	BB+/BB	0.548(0.034)	0.459(0.030)		
	KB/GL	0.769(0.026)	0.462(0.024)		
	HA/VW	1.176(0.044)	0.489(0.038)		

Table 9.1.2.2: Showing the population parameters generated by using OPMML module.

The first purpose of generating population parameters under different conditions is to make comparative study of how the moments differ within and between populations. The moments differing widely within a population indicates the need to examine the data while the differences of moments between populations may confirm logical expectations of a test designer.

The second purpose of estimating the population parameters conditioned on inclusion and exclusion of misfitting items and DIF items is to make a comparative study of moments across populations. Small differences in moments under different conditions for same population indicate negligible influence of misfitting items and DIF items. When the influence of misfitting items and DIF items is negligible, the population parameters obtained under the conditions 7, 8 and 9 are preferable option for use in making norm reference table as they have maximum number of items.

In case of UIBTERV English Reading Comprehension, the population parameters generated under the conditions 7, 8 and 9 are used in making global norm reference table.

9.1.3 Generating Global Norm by OPLAT Module

The population parameters obtained in section 9.1.2 under conditions 7, 8 and 9 are used for generating global norm by **OPLAT** module. The population parameters are computed into the **OPLAT.DEF** file. For details on **OPLAT.DEF** file, a reader may like to refer Verhelst, et al., (1995, p.93).

The screen file prepared by collapsing all booklets is *NERLSDDT.SCR*. The *NERLSDDT.SCR* file is used for running the **OPLAT** module by using *NERLSDD.PAR* file. The output file *NERLSDD.LAT* is generated by the **OPLAT** module. The *NERLSDD.LAT* file contains global norm as one of its contents. The global norm for UIBTERV English Reading Comprehension Test is provided in Appendix I.

In addition to the norm table, the moments of distributions are generated for each population.

Moments of Distribution

ALL	Mean(RMn)	St.Dev	RMSError	MAcc
Score	119.9(0.764)	23.638	4.597	0.962
Theta	0.846	0.552	0.140	0.934

Boys	Mean(RMn)	St.Dev	RMSError	MAcc
Score	120.9(0.770)	24.094	4.537	0.965
Theta	0.880	0.578	0.147	0.933

Girls	Mean(RMn)	St.Dev	RMSError	MAcc
Score	119.4(0.761)	22.632	4.647	0.958
Theta	0.822	0.517	0.133	0.932

Unknown	Mean(RMn)	St.Dev	RMSError	MAcc
Score	122.4(0.779)	22.457	4.512	0.960
Theta	0.903	0.546	0.145	0.926

BB+/BB	Mean(RMn)	St.Dev	RMSError	MAcc
Score	107.1(0.682)	24.743	5.030	0.959
Theta	0.553	0.477	0.108	0.948

KB/GL	Mean(RMn)	St.Dev	RMSError	MAcc
Score	117.9(0.751)	22.094	4.728	0.954
Theta	0.777	0.487	0.125	0.932

HA/VW	Mean(RMn)	St.Dev	RMSError	MAcc
Score	133.1(0.848)	16.916	4.010	0.944
Theta	1.194	0.536	0.183	0.877

9.1.4 Interpreting Global Norm Table and Moments of Distributions

Interpretation of global norm is quite straight forward. The score column contains scores ranging from minimum possible score to maximum possible score of a test. The theta column contains the theta values with which an examinee can get the scores corresponding to them. The mean column contains the mean of the theta values and the St.Dev. column contains the standard deviations. The standard deviation provides information about the distance of theta values from the mean theta value. The score distribution section contains percentile rank scores for different populations of examinees. From the

global norm, it can be seen that there are no examinees scoring 37 or less across different populations. The test scores start providing information about examinees only from the scores of 38 and above.

The moments of distribution has mean scores, mean theta values, mean of the estimated p-values (RMn), standard deviations (St.Dev), root mean square error (RMSE) and measure of accuracy (MAcc). Equations for RMSE and MAcc are given in chapter 7 in section 7.7.0. The moments indicate the performance of different groups of the examinees in terms of their mean scores. Moments are also useful for making confidence intervals.

For instance, the confidence interval for mean score of the ALL can be calculated by using equation 2.5.0.2 in chapter 2 and state as being between 110.106 and 128.674 at 95% confidence interval.

The global norm consists of 157 items and in real life situation a test with 157 items is not practicable unless it is a purely speed test. However, global norm can provide information to a test designer about how items function across populations and consequently help him or her in designing an effective test.

9.2.0 Item Information

Item information is defined in chapter 7 by equation 7.1.0.1. The OPLM generates item information along with global norm. The item information values and item p-values are used for selecting the items for a test. The item information and item p-values generated by OPLM for UIBTERV English Reading Comprehension Test are shown in table 9.2.0.1.

Item Nr.	Label	P-values				Item Information			
		ALL	BB+/BB	KB/GL	HA/VO	ALL	BB+/BB	KB/GL	HA/VO
8	50060_Le	0.657	0.55	0.64	0.78	0.74	0.84	0.79	0.60
9	40016_Le	0.806	0.73	0.80	0.89	0.54	0.69	0.57	0.36
10	40019_Le	0.822	0.72	0.81	0.92	0.99	1.37	1.08	0.57
11	40012_Le	0.37	0.31	0.35	0.45	0.22	0.20	0.22	0.23
12	40022_Le	0.965	0.95	0.96	0.98	0.13	0.19	0.13	0.07
21	30190_Le	0.87	0.81	0.86	0.93	0.40	0.55	0.43	0.25
22	30209_Le	0.833	0.79	0.83	0.88	0.13	0.16	0.14	0.11
23	80014_Le	0.661	0.55	0.64	0.78	0.74	0.84	0.79	0.60
24	80018_Le	0.806	0.73	0.80	0.89	0.54	0.69	0.57	0.36
25	80019_Le	0.765	0.68	0.75	0.86	0.61	0.75	0.65	0.43
31	40031_Le	0.957	0.93	0.96	0.98	0.16	0.23	0.16	0.08
32	40049_Le	0.899	0.85	0.90	0.95	0.33	0.46	0.35	0.19
33	40058_Le	0.942	0.91	0.94	0.97	0.20	0.30	0.21	0.11
34	40062_Le	0.954	0.93	0.95	0.98	0.17	0.25	0.17	0.09
35	40080_Le	0.855	0.77	0.85	0.94	0.86	1.23	0.93	0.46
41	40001_Le	0.946	0.91	0.95	0.98	0.39	0.64	0.41	0.17
43	40003_Le	0.967	0.95	0.97	0.98	0.12	0.18	0.13	0.06
44	40006_Le	0.982	0.97	0.98	0.99	0.07	0.11	0.07	0.04
45	40007_Le	0.692	0.59	0.67	0.80	0.71	0.82	0.75	0.55
51	80004_Le	0.873	0.80	0.87	0.95	0.78	1.14	0.84	0.40
52	80006_Le	0.61	0.54	0.60	0.68	0.22	0.24	0.23	0.21
53	80008_Le	0.805	0.70	0.79	0.91	1.05	1.43	1.15	0.63
54	80005_Le	0.787	0.70	0.78	0.87	0.57	0.72	0.61	0.40
55	80013_Le	0.845	0.76	0.84	0.93	0.90	1.28	0.98	0.50
60	40009_Le	0.804	0.72	0.79	0.89	0.54	0.69	0.58	0.37
61	40035_Le	0.874	0.82	0.87	0.93	0.39	0.53	0.41	0.24
62	40050_Le	0.828	0.75	0.82	0.90	0.49	0.65	0.53	0.33
63	40066_Le	0.558	0.44	0.53	0.69	0.81	0.84	0.84	0.72

Item Nr.	Label	P-values				Item Information			
		ALL	BB+/BB	KB/GL	HA/VO	ALL	BB+/BB	KB/GL	HA/VO
64	50062_Le	0.867	0.81	0.86	0.93	0.41	0.55	0.43	0.25
75	40011_Le	0.796	0.71	0.78	0.88	0.56	0.71	0.59	0.38
76	40015_Le	0.933	0.90	0.93	0.97	0.23	0.33	0.24	0.13
77	40023_Le	0.958	0.93	0.96	0.98	0.32	0.52	0.33	0.13
78	40025_Le	0.976	0.96	0.98	0.99	0.09	0.14	0.09	0.05
79	40028_Le	0.982	0.97	0.98	0.99	0.07	0.11	0.07	0.04
85	80026_Le	0.774	0.69	0.76	0.86	0.59	0.74	0.64	0.42
86	80028_Le	0.704	0.60	0.69	0.81	0.69	0.82	0.74	0.53
87	80029_Le	0.867	0.81	0.86	0.93	0.41	0.55	0.43	0.25
88	80037_Le	0.763	0.64	0.75	0.88	1.18	1.54	1.30	0.76
89	80027_Le	0.566	0.45	0.54	0.70	0.80	0.84	0.84	0.71
95	40029_Le	0.945	0.91	0.94	0.98	0.40	0.65	0.42	0.17
96	40030_Le	0.967	0.94	0.97	0.99	0.25	0.42	0.26	0.10
97	50056_Le	0.825	0.73	0.82	0.92	0.98	1.36	1.07	0.56
98	50075_Le	0.577	0.46	0.55	0.71	0.80	0.84	0.84	0.70
99	40076_Le	0.952	0.93	0.95	0.98	0.17	0.25	0.18	0.09
110	40018_Le	0.919	0.88	0.92	0.96	0.27	0.39	0.29	0.16
111	40010_Le	0.65	0.54	0.63	0.77	0.75	0.84	0.80	0.61
112	40021_Le	0.834	0.76	0.83	0.91	0.48	0.63	0.51	0.31
113	40032_Le	0.741	0.69	0.73	0.80	0.18	0.21	0.19	0.15
114	40036_Le	0.896	0.85	0.89	0.94	0.34	0.47	0.36	0.20
120	30078_Le	0.92	0.87	0.92	0.97	0.54	0.85	0.57	0.25
121	30183_Le	0.819	0.72	0.81	0.92	1.00	1.38	1.09	0.58
122	30181_Le	0.773	0.68	0.76	0.86	0.60	0.74	0.64	0.42
123	80023_Le	0.646	0.50	0.62	0.80	1.45	1.65	1.57	1.10
124	80021_Le	0.739	0.64	0.72	0.84	0.65	0.79	0.69	0.48
130	40046_Le	0.862	0.83	0.86	0.90	0.12	0.14	0.12	0.09
131	40048_Le	0.605	0.49	0.58	0.73	0.78	0.85	0.83	0.67
133	50038_Le	0.855	0.77	0.85	0.94	0.85	1.23	0.93	0.46
134	50077_Le	0.592	0.47	0.57	0.72	0.79	0.85	0.83	0.69
145	40004_Le	0.926	0.89	0.92	0.96	0.25	0.36	0.27	0.14
147	40017_Le	0.986	0.98	0.99	0.99	0.05	0.08	0.05	0.03
148	40033_Le	0.932	0.90	0.93	0.97	0.23	0.34	0.25	0.13
149	40056_Le	0.803	0.76	0.80	0.85	0.15	0.18	0.16	0.12
157	30162_Le	0.839	0.77	0.83	0.91	0.47	0.62	0.50	0.30
158	30220_Le	0.863	0.83	0.86	0.90	0.12	0.14	0.12	0.09
159	30128_Le	0.839	0.80	0.83	0.88	0.13	0.16	0.14	0.10
160	30133_Le	0.862	0.77	0.86	0.95	1.27	1.94	1.40	0.61
161	30150_Le	0.614	0.50	0.59	0.74	0.78	0.85	0.82	0.66
169	50001_Le	0.622	0.56	0.61	0.70	0.22	0.24	0.23	0.20
170	50078_Le	0.767	0.68	0.75	0.86	0.61	0.75	0.65	0.43
171	50079_Le	0.775	0.69	0.76	0.87	0.59	0.74	0.63	0.42
172	50052_Le	0.851	0.79	0.84	0.92	0.44	0.59	0.47	0.28
173	50071_Le	0.758	0.64	0.74	0.88	1.20	1.55	1.31	0.78
180	40054_Le	0.85	0.78	0.84	0.92	0.45	0.60	0.48	0.28
181	40057_Le	0.795	0.69	0.78	0.90	1.08	1.46	1.18	0.66
182	40072_Le	0.687	0.58	0.67	0.80	0.71	0.83	0.76	0.56
183	50024_Le	0.838	0.77	0.83	0.91	0.47	0.63	0.51	0.31
189	30250_Le	0.822	0.75	0.81	0.90	0.51	0.66	0.54	0.34

Item Nr.	Label	P-values				Item Information			
		ALL	BB+/BB	KB/GL	HA/VO	ALL	BB+/BB	KB/GL	HA/VO
190	30249_Le	0.766	0.68	0.75	0.86	0.61	0.75	0.65	0.43
191	30330_Le	0.861	0.78	0.86	0.94	0.83	1.20	0.90	0.44
192	30394_Le	0.806	0.73	0.80	0.89	0.54	0.69	0.57	0.36
193	30270_Le	0.657	0.55	0.64	0.78	0.74	0.84	0.79	0.60
196	50031_Le	0.772	0.72	0.76	0.83	0.17	0.19	0.17	0.14
198	50068_Le	0.858	0.77	0.85	0.94	0.84	1.22	0.91	0.45
199	50063_Le	0.747	0.62	0.73	0.87	1.23	1.57	1.35	0.81
200	50047_Le	0.946	0.92	0.94	0.97	0.19	0.28	0.20	0.10
212	40053_Le	0.693	0.59	0.68	0.81	0.71	0.82	0.75	0.55
213	40064_Le	0.617	0.55	0.60	0.69	0.22	0.24	0.23	0.20
214	40077_Le	0.874	0.84	0.87	0.91	0.11	0.13	0.11	0.08
216	50030_Le	0.644	0.53	0.62	0.77	0.76	0.84	0.80	0.62
222	30246_Le	0.833	0.72	0.82	0.93	1.45	2.12	1.61	0.75
223	30245_Le	0.829	0.76	0.82	0.90	0.49	0.64	0.52	0.32
224	30226_Le	0.687	0.58	0.67	0.80	0.71	0.83	0.76	0.56
225	30227_Le	0.771	0.68	0.76	0.86	0.60	0.75	0.64	0.42
226	30247_Le	0.792	0.68	0.78	0.90	1.09	1.47	1.20	0.67
232	50057_Le	0.392	0.25	0.35	0.57	1.50	1.28	1.52	1.58
233	50040_Le	0.835	0.74	0.83	0.92	0.94	1.32	1.02	0.53
234	50041_Le	0.922	0.88	0.92	0.96	0.26	0.38	0.28	0.15
235	50042_Le	0.749	0.65	0.73	0.85	0.63	0.78	0.68	0.46
236	50043_Le	0.832	0.74	0.82	0.92	0.95	1.33	1.04	0.54
250	40045_Le	0.488	0.37	0.46	0.63	0.82	0.79	0.84	0.78
251	40055_Le	0.714	0.61	0.70	0.82	0.68	0.81	0.73	0.52
252	40068_Le	0.8	0.69	0.79	0.90	1.07	1.44	1.17	0.64
253	40070_Le	0.922	0.87	0.92	0.97	0.53	0.83	0.57	0.25
254	40071_Le	0.703	0.65	0.69	0.77	0.20	0.22	0.20	0.17
262	30280_Le	0.687	0.58	0.67	0.80	0.71	0.83	0.76	0.56
263	30196_Le	0.533	0.41	0.51	0.67	0.81	0.82	0.85	0.75
264	30125_Le	0.808	0.76	0.80	0.86	0.15	0.17	0.16	0.12
265	30098_Le	0.727	0.63	0.71	0.83	0.66	0.80	0.71	0.50
266	30053_Le	0.592	0.48	0.57	0.72	0.79	0.85	0.83	0.68
271	40073_Le	0.801	0.76	0.79	0.85	0.15	0.18	0.16	0.12
272	50070_Le	0.834	0.74	0.83	0.92	0.94	1.32	1.03	0.53
273	50039_Le	0.895	0.85	0.89	0.94	0.34	0.47	0.36	0.20
274	40051_Le	0.821	0.78	0.81	0.87	0.14	0.17	0.15	0.11
275	50044_Le	0.77	0.68	0.76	0.86	0.60	0.75	0.64	0.43
276	40008_Le	0.779	0.73	0.77	0.83	0.17	0.19	0.17	0.14
277	40034_Le	0.754	0.63	0.74	0.88	1.21	1.56	1.33	0.79
278	40038_Le	0.847	0.78	0.84	0.91	0.45	0.60	0.48	0.29
279	40044_Le	0.489	0.42	0.47	0.57	0.24	0.23	0.24	0.23
280	30325_Le	0.797	0.75	0.79	0.85	0.16	0.18	0.16	0.13
281	30342_Le	0.715	0.66	0.70	0.78	0.19	0.22	0.20	0.17
282	30462_Le	0.678	0.54	0.65	0.82	1.39	1.65	1.51	1.01
283	30172_Le	0.738	0.64	0.72	0.84	0.65	0.79	0.69	0.48
284	30130_Le	0.679	0.57	0.66	0.79	0.72	0.83	0.77	0.57
291	40059_Le	0.884	0.83	0.88	0.94	0.37	0.51	0.39	0.22
292	40075_Le	0.884	0.83	0.88	0.94	0.37	0.51	0.39	0.22
293	50080_Le	0.754	0.63	0.74	0.88	1.21	1.56	1.33	0.79

Item Nr.	Label	P-values				Item Information			
		ALL	BB+/BB	KB/GL	HA/VO	ALL	BB+/BB	KB/GL	HA/VO
294	50059_Le	0.801	0.72	0.79	0.88	0.55	0.70	0.58	0.37
305	40037_Le	0.731	0.63	0.72	0.83	0.66	0.79	0.70	0.49
306	40039_Le	0.248	0.16	0.22	0.37	0.63	0.48	0.60	0.78
307	40042_Le	0.692	0.59	0.67	0.80	0.71	0.82	0.75	0.55
308	50066_Le	0.845	0.76	0.84	0.93	0.90	1.28	0.98	0.50
314	17035_Le	0.607	0.43	0.57	0.79	2.27	2.51	2.49	1.76
315	21042_Le	0.785	0.70	0.77	0.87	0.58	0.73	0.62	0.40
316	21055_Le	0.687	0.63	0.68	0.75	0.20	0.22	0.21	0.18
317	30081_Le	0.551	0.43	0.52	0.69	0.81	0.83	0.84	0.73
318	30140_Le	0.529	0.38	0.49	0.70	1.55	1.57	1.65	1.38
323	50067_Le	0.485	0.33	0.45	0.66	1.56	1.49	1.63	1.46
324	50037_Le	0.551	0.43	0.52	0.69	0.81	0.83	0.84	0.73
325	40074_Le	0.562	0.44	0.54	0.70	0.80	0.84	0.84	0.72
326	50050_Le	0.768	0.65	0.75	0.88	1.17	1.53	1.28	0.74
327	50054_Le	0.631	0.52	0.61	0.76	0.77	0.85	0.81	0.64
330	40040_Le	0.879	0.82	0.87	0.93	0.38	0.52	0.40	0.23
331	40043_Le	0.916	0.87	0.91	0.96	0.28	0.40	0.30	0.16
332	40047_Le	0.754	0.66	0.74	0.85	0.63	0.77	0.67	0.45
333	40060_Le	0.556	0.49	0.54	0.63	0.23	0.24	0.24	0.22
334	40065_Le	0.754	0.66	0.74	0.85	0.63	0.77	0.67	0.45
342	01115_Le	0.492	0.34	0.45	0.67	1.56	1.50	1.64	1.45
343	04038_Le	0.798	0.72	0.79	0.88	0.55	0.70	0.59	0.38
344	30060_Le	0.649	0.54	0.63	0.77	0.75	0.84	0.80	0.61
345	30142_Le	0.548	0.43	0.52	0.68	0.81	0.83	0.84	0.73
346	30168_Le	0.816	0.77	0.81	0.86	0.14	0.17	0.15	0.12
353	50026_Le	0.602	0.49	0.58	0.73	0.79	0.85	0.83	0.67
354	50055_Le	0.637	0.52	0.62	0.76	0.76	0.85	0.81	0.63
355	50045_Le	0.406	0.29	0.38	0.55	0.79	0.71	0.80	0.83
356	50035_Le	0.862	0.80	0.86	0.92	0.42	0.57	0.45	0.26
357	50034_Le	0.719	0.59	0.70	0.85	1.30	1.61	1.42	0.89
358	=E279_Le	0.573	0.51	0.56	0.65	0.23	0.24	0.24	0.22

Table 9.2.0.1: Shows item information and item p-values.

The usage of the table 9.2.0.1 is explained in chapter 10.

9.3.0 Summary

Percentile ranks are used as normative scores in UIBTERV English Reading Comprehension Test.

The estimation of population parameters have to be done several times based on different conditions applied to the data. Population parameters will enable a test designer to make comparative studies of scores both between and within populations. The information obtained from the comparative study of the moments may aid a test designer to (a) investigate the response data, (b) confirm his/her logical expectations and (c) investigate the effects of items identified as problem items.

9.4.0 References

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *One-Parameter Logistic Model*, Cito, Arnhem, the Netherlands.

Chapter 10

Making English Reading Comprehension Entrance Test

10.0.0 Introduction

To make tests from the pool of items with known parameters requires defining the goals of the tests, the use of test information function, to generate test specific norms and comparative study of the moments across different groups of examinees. However, the comparative study of the moments across different groups of examinees may not be always relevant if the examinees are from a homogeneous group.

In this chapter, the method used for selecting items from the pool of calibrated items for making UIBTERV English Reading Comprehension Entrance Test will be described.

10.1.0 Assembling Tests from Item Pool

To assemble tests from item pool, the goals of the tests will have to be defined. Precise goals of the tests will enable the test designers to specify the range of abilities that the tests are supposed to measure and accordingly assemble the items which provide maximum information around that range of abilities.

10.1.1 Specifying Ability Ranges

In case of UIBTERV English Reading Comprehension Entrance Test, the goal of the test is to measure the reading competence of the students at the beginning of the first year of the Lower Secondary School.

The reading comprehension competence is measured across different levels of examinees. Therefore the range of abilities that the English Reading Comprehension Entrance Test has to measure are verbally described as (a) easy and average easy range of abilities, (b) average easy and average difficult range of abilities and (c) average difficult and difficult range of abilities.

The estimated p-values of the items from an output file **.LAT* generated by **OPLAT** module is used in selecting the items for making tests to measure the three types of abilities. The output file is the same file that is used to generate global norm reference table, i.e., *NERLSDD.LAT*.

Based on the estimated p-values, the items are categorized into four groups, viz., (a) easy items, (b) average easy items, (c) average difficult items and (d) difficult items. Table 10.1.1.1 shows the items in each category with their estimated proportion correct values. The maximum p-value of the items is 0.896 and the minimum p-value of the items is 0.492. The items with p-values more than 0.896 are not selected for the tests as they are unlikely to be discriminative due to possible ceiling effect. The items with p-values less than 0.492 are not selected for the tests as they are unlikely to be discriminative due to possible floor effect.

Easy Item		Average Easy Item		Average Difficult Item		Difficult Item	
Item No.	P-Value	Item No.	P-Value	Item No.	P-Value	Item No.	P-Value
173	0.758	265	0.727	123	0.646	342	0.492
181	0.795	332	0.754	224	0.687	318	0.529
252	0.8	190	0.766	305	0.731	345	0.548
192	0.806	326	0.768	283	0.738	324	0.551
62	0.828	225	0.771	199	0.747	89	0.566

Easy Item		Average Easy Item		Average Difficult Item		Difficult Item	
Item No.	P-Value	Item No.	P-Value	Item No.	P-Value	Item No.	P-Value
159	0.839	122	0.773	277	0.754	314	0.607
308	0.845	54	0.787	293	0.754	216	0.644
133	0.855	343	0.798	334	0.754	123	0.646
198	0.858	60	0.804	88	0.763	8	0.657
191	0.861	53	0.805	25	0.765	23	0.661
87	0.867	9	0.806	170	0.767	284	0.679
21	0.87	24	0.806	171	0.775	212	0.693
330	0.879	10	0.822	276	0.779	86	0.704
292	0.884	97	0.825	315	0.785	251	0.714
114	0.896	278	0.847	346	0.816	357	0.719

Table 10.1.1.1: Items with p-values in four categories.

The items in groups (a) and (b) are mixed together to make a test comprising of easy and average easy items, the items from groups (b) and (c) are mixed together to make a test comprising of average easy and average difficult items and the items from groups (c) and (d) are mixed together to make a test comprising of average difficult items and difficult items. The tests are shown as T_1, T_2 and T_3 in table 10.1.1.2.

$$T_1 = \left\{ \begin{matrix} 9,10,21,24,53,54,60,62,87,97,114,122,133,159,173,181,190,191,192,198, \\ 225,265,278,292,308,326,330,332,343 \end{matrix} \right\}$$

$$T_2 = \left\{ \begin{matrix} 9,10,24,25,53,54,60,88,97,122,123,170,171,190,199,224,225,265,276, \\ 277,278,283,293,305,315,326,332,334,343,346 \end{matrix} \right\}$$

$$T_3 = \left\{ \begin{matrix} 8,23,25,86,88,89,123,170,171,199,212,216,224,251,276,277,283,284, \\ 293,305,314,315,318,324,334,342,345,343,355,357 \end{matrix} \right\}$$

Table 10.1.1.2: Shows the items in three different tests for three groups of examinees.

Notice that the tests have common items called anchor items. The anchor items enable the test makers to place the tests on a common scale. When the tests have a common scale, the performances of students in different tests are comparable. The anchor items are identified in table 10.1.1.3

$$T_1 \cap T_2 = \{9,10,53,54,60,97,122,190,225,265,278,326,332,343\}$$

$$T_2 \cap T_3 = \{25,88,123,170,171,199,224,276,277,283,293,305,315,334,346\}$$

Table 10.1.1.3: Shows the anchor items.

10.2.0 Making Norm Reference Table for Tests

The test specific norm reference table or local norm reference tables are produced by following the steps similar to those used in generating global norm reference table. The files used for making local norms are *NERSXT3R.SCR* and *NERDDSX.PAR*. The out put file is *NERSXT3R.LAT*.

The population parameters supplied to *OPLAT.DEF* file are shown in table 10.2.0.

Population	Mu	Sigma
1 BB+/BB	0.548	0.459
2 KB/GL	0.769	0.462
3 HA/VW	1.176	0.489

Table 10.2.0.1: Shows the population parameters supplied to *OPLAT.DEF* file.

The screen file *NERSXT3R.SCR* consists of three booklets. The items from T_1 are transferred to booklet one, the items from T_2 are transferred to booklet two and the items from T_3 are transferred to booklet three. Each booklet, therefore, produces a test with complete information that is available in conventional **OPLAT** output file.

The three tests are graphically displayed in figure10.2.0.2. The test with average difficult and difficult items consistently requires more ability than the other two tests to get a common score and the test with average easy and average difficult items consistently requires more ability than the test with easy and average easy items to get a common score.

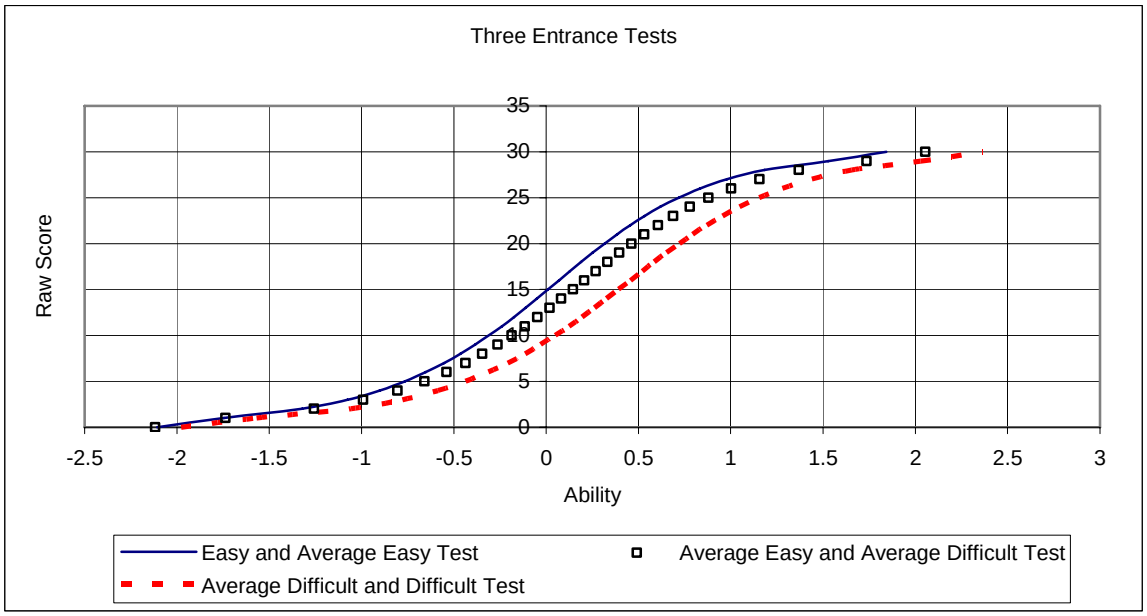

Figure 10.2.0.2: Three versions of English Reading Comprehension test.

The examinees taking different tests having the same raw score can be differentiated in terms of the differences in their abilities. Table 10.2.0.3 shows the data used for generating figure 10.2.0.2.

Score	Test1 (Ability)	Test2 (Ability)	Test3 (Ability)
0	-2.108	-2.117	-1.977
1	-1.746	-1.737	-1.569
2	-1.321	-1.257	-1.058
3	-1.077	-0.991	-0.779
4	-0.903	-0.805	-0.586
5	-0.766	-0.66	-0.437
6	-0.652	-0.54	-0.314
7	-0.553	-0.437	-0.208
8	-0.465	-0.346	-0.114
9	-0.385	-0.263	-0.029
10	-0.311	-0.186	0.05
11	-0.241	-0.115	0.124
12	-0.175	-0.047	0.194
13	-0.112	0.019	0.262
14	-0.051	0.082	0.327
15	0.01	0.145	0.392
16	0.07	0.207	0.456
17	0.129	0.269	0.521
18	0.19	0.331	0.586
19	0.252	0.396	0.653
20	0.316	0.462	0.722
21	0.383	0.532	0.794
22	0.455	0.607	0.871
23	0.533	0.688	0.955
24	0.619	0.778	1.048
25	0.717	0.88	1.154
26	0.834	1.002	1.28
27	0.98	1.156	1.439
28	1.183	1.369	1.657
29	1.527	1.736	2.028
30	1.842	2.054	2.365

Table 10.2.0.3: Data used for figure 10.2.0.2

For instance, an examinee with 10 scores in Test1 is approximately equal to an examinee with 8 scores in Test2 and an examinee with 7 scores in Test3.

10.3.0 Norm Reference Table for 3 Tests

The norm reference tables for the 3 tests are shown in tables 10.3.1, 10.3.2 and 10.3.3.

Conditional Distribution				Score Distribution		
Score	Theta	Mean	St.Dev	BB+/BB	KB/GL	HA/VO
0	-2.108	-1.897	0.267	0	0	0
1	-1.746	-1.74	0.337	0	0	0
2	-1.321	-1.419	0.387	0	0	0
3	-1.077	-1.167	0.365	0	0	0
4	-0.903	-0.972	0.324	0	0	0
5	-0.766	-0.818	0.285	1	0	0
6	-0.652	-0.691	0.253	1	0	0
7	-0.553	-0.584	0.23	2	1	0
8	-0.465	-0.489	0.212	2	1	0
9	-0.385	-0.404	0.199	3	1	0
10	-0.311	-0.326	0.189	5	2	0
11	-0.241	-0.254	0.181	6	2	0
12	-0.175	-0.185	0.175	8	3	1
13	-0.112	-0.119	0.171	10	4	1
14	-0.051	-0.055	0.168	12	5	1
15	0.01	0.008	0.167	15	7	1
16	0.07	0.07	0.166	18	9	2
17	0.129	0.132	0.167	21	11	2
18	0.19	0.195	0.168	25	13	3
19	0.252	0.26	0.172	29	16	4
20	0.316	0.328	0.177	34	20	6
21	0.383	0.399	0.184	40	24	7
22	0.455	0.475	0.193	46	29	10
23	0.533	0.558	0.207	52	35	13
24	0.619	0.653	0.226	59	42	17
25	0.717	0.762	0.251	67	50	23
26	0.834	0.893	0.282	75	60	31
27	0.98	1.057	0.314	83	71	42
28	1.183	1.266	0.331	91	82	58
29	1.527	1.531	0.29	97	93	80
30	1.842	1.672	0.227	100	100	100

Table 10.2.1.1: Norm Reference table for Easy and Average Test.

Moments of Distribution				
BB+/BB	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	22.0 (0.733)	5.713	2.188	0.853
Theta	0.516	0.497	0.242	0.782
KB/GL	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	24.2 (0.807)	4.858	1.98	0.834
Theta	0.672	0.499	0.267	0.75
HA/VO	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	27.0 (0.899)	3.35	1.541	0.788
Theta	0.785	0.578	0.307	0.717

Conditional Distribution				Score Distribution		
Score	Theta	Mean	St.Dev	BB+/BB	KB/GL	HA/VO
0	-2.117	-1.879	0.295	0	0	0
1	-1.737	-1.721	0.366	0	0	0
2	-1.257	-1.368	0.423	0	0	0
3	-0.991	-1.093	0.398	0	0	0
4	-0.805	-0.883	0.35	1	0	0
5	-0.66	-0.718	0.305	1	0	0
6	-0.54	-0.584	0.27	2	1	0
7	-0.437	-0.471	0.243	3	1	0
8	-0.346	-0.372	0.223	4	2	0
9	-0.263	-0.284	0.208	6	2	0
10	-0.186	-0.203	0.197	8	3	1
11	-0.115	-0.128	0.189	10	4	1
12	-0.047	-0.056	0.183	13	6	1
13	0.019	0.012	0.178	15	7	1
14	0.082	0.078	0.175	19	9	2
15	0.145	0.143	0.174	22	11	3
16	0.207	0.207	0.173	26	14	3
17	0.269	0.272	0.174	31	17	5
18	0.331	0.337	0.176	35	20	6
19	0.396	0.404	0.179	40	24	7
20	0.462	0.474	0.184	46	29	10
21	0.532	0.548	0.192	52	34	12
22	0.607	0.628	0.203	58	40	16
23	0.688	0.715	0.217	64	47	20
24	0.778	0.813	0.238	71	54	26
25	0.88	0.928	0.264	78	63	33
26	1.002	1.066	0.298	84	72	43
27	1.156	1.238	0.331	90	81	55
28	1.369	1.457	0.347	95	89	70
29	1.736	1.734	0.302	98	96	87
30	2.054	1.873	0.238	100	100	100

Table 10.2.1.2: Norm Reference table for Average Easy and Average Difficult Test.

	Moments of Distribution			
BB+/BB	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	20.3(0.676)	6.049	2.319	0.853
Theta	0.541	0.51	0.236	0.791
KB/GL	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	22.7 (0.757)	5.349	2.148	0.839
Theta	0.732	0.506	0.261	0.758
HA/VO	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	26.0 (0.866)	3.92	1.736	0.804
Theta	0.947	0.547	0.305	0.72

Conditional Distribution				Score Distribution		
Score	Theta	Mean	St.Dev	BB+/BB	KB/GL	HA/VO
0	-1.977	-1.721	0.312	0	0	0
1	-1.569	-1.553	0.386	0	0	0
2	-1.058	-1.176	0.442	1	0	0
3	-0.779	-0.885	0.412	1	0	0
4	-0.586	-0.666	0.359	2	1	0
5	-0.437	-0.496	0.311	4	1	0
6	-0.314	-0.358	0.273	6	2	0
7	-0.208	-0.241	0.245	8	3	1
8	-0.114	-0.14	0.225	11	5	1
9	-0.029	-0.049	0.21	14	6	1
10	0.05	0.034	0.199	17	8	2
11	0.124	0.111	0.191	21	11	3
12	0.194	0.185	0.184	26	14	3
13	0.262	0.255	0.18	30	17	4
14	0.327	0.323	0.177	35	20	6
15	0.392	0.39	0.175	40	24	7
16	0.456	0.457	0.174	45	28	9
17	0.521	0.524	0.175	50	33	11
18	0.586	0.591	0.177	56	38	14
19	0.653	0.661	0.18	61	43	17
20	0.722	0.733	0.186	66	49	21
21	0.794	0.81	0.193	72	55	26
22	0.871	0.892	0.204	77	61	31
23	0.955	0.982	0.218	82	68	37
24	1.048	1.083	0.239	86	74	44
25	1.154	1.201	0.266	90	81	53
26	1.28	1.342	0.3	94	86	62
27	1.439	1.519	0.335	97	92	73
28	1.657	1.745	0.353	98	96	84
29	2.028	2.031	0.311	100	99	94
30	2.365	2.183	0.243	100	100	100

Table 10.2.1.2: Norm Reference table for Average Difficult and Difficult Test.

	Moments of Distribution			
BB+/BB	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	17.1 (0.570)	6.373	2.388	0.86
Theta	0.547	0.52	0.217	0.817
KB/GL	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	19.8 (0.661)	6.023	2.297	0.855
Theta	0.766	0.515	0.232	0.798
HA/VO	Mean (RMn)	St.Dev	RMS Error	MAcc
Score	23.9 (0.798)	4.906	1.977	0.838
Theta	1.093	0.527	0.28	0.753

10.4.0 Summary

Well defined and precise goals of a test will enable the test designers to specify the ability ranges to be measured. UIBTERV English Reading Comprehension Entrance Test measures the ability in (a) easy and average range, (b) average easy and average difficult range and (c) average difficult and difficult range.

The items for the ability ranges are selected by using the estimated p-values of the items from the item pool. The item pool consists of calibrated items.

The items having p-values more than 0.896 are not used in the tests as they are likely to induce ceiling effect in the test. The items having p-values less than 0.492 are not used in the tests as they are likely to induce floor effect in the tests.

Different tests with anchor items are comparable and as a result, examinees from different tests can be compared with respect to their scores and abilities.

Chapter 11

Item Banking and Linking New Items to Old Bank

11.0.0 Introduction

Development of test specifications, item writing, field testing the items and calibrating the items before selecting the good items for a final test every time a test is administered is a tedious and expensive process. To avoid the repetition of the same process every time a test is administered, item banking is becoming more and more popular among the testing centers. Item bank is a data base which contains items matched to objectives, skills and curriculum content areas. An item bank has to be constantly replenished with new items to fill the void created by the use of the old items in a test. The method of depositing new items in an item bank involves a special technique called item linking.

In this chapter, item banking and item linking will be described as applied in banking and linking items from UIBTERV English Reading Comprehension Entrance Test to the old item bank. The results of item parameters obtained by using the method used in OPLM and Mean and Sigma method of placing items on a common scale will be compared.

11.1.0 Item Bank and its Functions

Gronlund (1998, p.130) define item banks as the files of various suitable test items that are coded by subject area, instructional level, instructional objective measured and various pertinent item characteristics like item difficulty and item discriminating power. For an item to be eligible for banking, it must be free from aberrant functions, bias and have to be accompanied by its parameters.

Item bank has a huge potential of easing and improving test construction process. A test developer can make a test to measure objectives of interest with a desired number of items by using an item bank. A test developer can construct different tests with predictable characteristics by using the items from an item bank and compare the performance of examinees who sat for different tests.

Item bank provides information to the curriculum developers and educational policy makers for deciding curriculum goals and objectives. Rudner (1998) notes that since the items describe individual tasks students are capable or incapable of doing, the location of the items on a calibrated scale allows one to identify the relative difficulty of particular tasks providing a way to discuss possible learning hierarchies and ways to better structure curriculum.

According to Ward (1994, p.35), an item bank can assist a test developer to (a) do item entry and storage (b) do item retrieval for reviewing items, formatting test forms, and editing and updating items and (c) maintain item history.

Item bank can be extended by depositing additional items by linking technique.

11.2.0 Building Item Bank

Although item bank has an enormous potential to ease and improve test construction process, it demands skills and professional expertise. The following steps are involved in building an item bank:

1) The goals and objectives of item bank have to be identified.

2) Appropriate people will have to be identified for developing items and performing item-content matching.

3) The items have to be field tested through different tests across wide range of abilities.

4) The items from different tests have to be calibrated on a common scale by using suitable IRT models.

5) Item bank data base has to be developed.

6) Item bank has to be replenished with new items.

7) Item bank users should possess a good knowledge of computer and other related computer packages used in item bank.

The steps 4 and 6 will be described further.

11.2.1 Calibrating Items from Different Tests on a Common Scale

The items in an item bank should have a common scale whether they are the items from same test or from different tests. The test design applied to UIBTERV English Reading Comprehension Test that makes the calibration of the items from different tests possible on a common scale is the anchor test design.

An anchor test design uses anchor items to link test X to test Y. Supposing that N_X examinees take test X which has n_x items with n_a anchor items and N_y examinees take test Y which has n_y items with n_a anchor items, then the test X and the test Y are linked by the anchor items and accordingly their relationship is established through the anchor items (Hambleton and Swaminathan, 1985, pp.211-212).

The anchor item design is used in field testing the potential test items for UIBTERV English Reading Comprehension Entrance Test. Thirteen test booklets consisting of 70 items each are linked by anchor items. The OPLM is used to calibrate the response data from the 13 test booklets in a single computer run. In this way the item parameters from 13 test booklets are placed on a common scale.

The items from 13 booklets can be used for making tests without concerning about the scaling problem. The items not used in the tests are too expensive to be discarded. They have to be deposited in an item bank.

11.2.2 Replenishing Item Bank

To replenish an item bank means depositing new items in it. One important property of an item bank is that the items in it should be matched to the content and objectives of a curriculum and the items measuring the same content and objectives of a curriculum should be on a common scale. This means that the new items will have to have the same scale as that of the other items in an item bank or vice versa.

To place the new items and the items in an item bank on a common scale, the principle of an anchor item design is used for extending the item bank of English Reading Comprehension by adding new items. Forty items from the old item bank are used as anchor items in field testing the English Reading Comprehension items. The number of items involved in field testing is 157 items inclusive of the anchor items. The items are written in 13 test booklets linked by other anchor items. It may be noted that the forty items from the item bank were not necessarily the anchor items across all booklets. The results from the pilot tests are analyzed by using OPLM. OPLM was also used in calibrating the items in the item bank.

To place the new items and the items in the item bank on a common scale, scaling constants have to be determined. Hambleton and Swaminathan (1985, p.205) present various ways of determining the scaling constants. In this thesis, Mean and Sigma Method and the method used in OPLM will be

discussed with reference to linking English Reading Comprehension items to other items in the item bank.

11.2.3 Mean and Sigma Method of Determing Scaling Constants

By assuming that the parameters of the anchor items as obtained from the field tests data and as they are in the item bank are linearly related (Hambleton et al., 1985, pp.131-132), linear relationships can be established as

$$b_{ya} = \alpha b_{xa} + \beta \qquad (11.2.2.1)$$

$$a_{ya} = \frac{a_{xa}}{\alpha} \qquad (11.2.2.2)$$

In equations 11.2.2.1 and 11.2.2.2, b_{ya} an b_{xa} are the item difficulties of anchor items in the item bank and the newly calibrated items, a_{ya} and a_{xa} are the item discriminations of anchor items in the item bank and the newly calibrated items and α and β are the scaling constants for item discrimination and item difficulty.

Further, the means and the standard deviations of the anchor item parameters have linear relationships as

$$\bar{b}_{ya} = \alpha \bar{b}_{xa} + \beta \qquad (11.2.2.3)$$

$$s_{ya} = \alpha s_{xa} \qquad (11.2.2.4)$$

where $\bar{b}_{ya}, \bar{b}_{xa}, s_{ya}$ and s_{xa} are the means and standard deviations of the parameters of anchor items in item bank and field tests respectively.

From equations 11.2.2.3 and 11.2.2.4, the scaling constants can be derived as

$$\alpha = \frac{s_{ya}}{s_{xa}} \qquad (11.2.2.5)$$

$$\beta = \bar{b}_{ya} - \alpha \bar{b}_{xa} \qquad (11.2.2.6)$$

Once α and β are determined, the item parameter estimates for field tested items are placed on the same scale as the items in the item bank by using the relationships

$$b_y^* = \alpha b_x + \beta \qquad (11.2.2.7)$$

$$a_y^* = \frac{a_x}{\alpha} \qquad (11.2.2.8)$$

where b_y^* and a_y^* are the difficulty and discrimination values of items in field tests placed on the scale of items in the item bank.

When the parameters of the new items are adjusted to with the scaling constants, the parameters of the anchor items in the new item group will also change. To account for the difference in the parameters of the anchor items in the new item group and in the item bank, the parameters of the anchor items in

the new item group are averaged with the parameters of the anchor items in the item bank and the resulting parameters are assigned to the anchor items (Hambleton and Swaminathan, 1985, p. 137).
In table 11.2.2.1, the parameters of the anchor items in the item bank and field tests are displayed. As can be seen in the table 11.2.2.1, parameters are different. The apparent differences are usually attributed to sampling fluctuation.

Parameters of Anchor Item in Item Bank				Parameters of Anchor Item in English Reading Comprehension Test			
Item No	Item Label	a	b	Item No.	Item Label	a	b
342	01115 _Le	3	0.011	342	01115_Le	3	0.852
343	04038 _Le	2	-0.368	343	04038_Le	2	0.008
314	17035 _Le	3	0.018	314	17035_Le	4	0.653
315	21042 _Le	2	-0.276	315	21042_Le	2	0.055
316	21055 _Le	3	0.043	316	21055_Le	1	0.004
266	30053 _Le	5	-0.074	266	30053_Le	2	0.608
344	30060 _Le	5	-0.061	344	30060_Le	2	0.459
120	30078 _Le	4	-0.341	120	30078_Le	3	-0.271
317	30081 _Le	2	0.011	317	30081_Le	2	0.712
265	30098 _Le	3	-0.152	265	30098_Le	2	0.239
264	30125 _Le	2	-0.642	264	30125_Le	1	-0.681
159	30128 _Le	2	-0.714	159	30128_Le	1	-0.902
284	30130 _Le	4	-0.133	284	30130_Le	2	0.38
160	30133 _Le	5	-0.337	160	30133_Le	4	0.093
318	30140 _Le	4	0.042	318	30140_Le	3	0.78
345	30142 _Le	2	-0.068	345	30142_Le	2	0.718
161	30150 _Le	4	0.003	161	30150_Le	2	0.551
157	30162 _Le	3	-0.56	157	30162_Le	2	-0.154
346	30168 _Le	1	-1.087	346	30168_Le	1	-0.735
283	30172 _Le	3	-0.24	283	30172_Le	2	0.208
122	30181 _Le	3	-0.371	122	30181_Le	2	0.095
121	30183 _Le	4	-0.182	121	30183_Le	3	0.127
21	30190 _Le	2	-0.681	21	30190_Le	2	-0.295
263	30196 _Le	3	0.009	263	30196_Le	2	0.754
22	30209 _Le	2	-0.786	22	30209_Le	1	-0.858
158	30220 _Le	2	-0.682	158	30220_Le	1	-1.099
224	30226 _Le	4	-0.112	224	30226_Le	2	0.357
225	30227 _Le	4	-0.139	225	30227_Le	2	0.101
223	30245 _Le	2	-0.392	223	30245_Le	2	-0.113
222	30246 _Le	6	-0.344	222	30246_Le	4	0.178
226	30247 _Le	4	-0.284	226	30247_Le	3	0.206
190	30249 _Le	5	-0.114	190	30249_Le	2	0.118
189	30250 _Le	4	-0.333	189	30250_Le	2	-0.083
193	30270 _Le	4	-0.112	193	30270_Le	2	0.439
262	30280 _Le	5	-0.045	262	30280_Le	2	0.357
280	30325 _Le	1	-0.998	280	30325_Le	1	-0.612
191	30330 _Le	3	-0.464	191	30330_Le	3	-0.011
281	30342 _Le	1	-0.964	281	30342_Le	1	-0.14
192	30394 _Le	3	-0.52	192	30394_Le	2	-0.023
282	30462 _Le	3	-0.231	282	30462_Le	3	0.48

Table 11.2.2.1: Shows the parameters of anchor items.

To place the parameters of the anchor items obtained from the field testing of English Reading Comprehension items on the common scale as the other items in the item bank, the mean of the

parameters of the anchor items as they are in item bank and as they are calibrated after field testing are calculated based in equations 11.2.2.3 through 11.2.2.6. The procedure is shown in table 11.2.2.2.

Item No.	Label	Item Bank		Field Tests	
		Discrimination (a_{ya})	Difficulty (b_{ya})	Discrimination (a_{xa})	Difficulty (b_{xa})
342	01115 _Le	3	0.011	3	0.852
343	04038 _Le	2	-0.368	2	0.008
314	17035 _Le	3	0.018	4	0.653
315	21042 _Le	2	-0.276	2	0.055
316	21055 _Le	3	0.043	1	0.004
266	30053 _Le	5	-0.074	2	0.608
344	30060 _Le	5	-0.061	2	0.459
120	30078 _Le	4	-0.341	3	-0.271
317	30081 _Le	2	0.011	2	0.712
265	30098 _Le	3	-0.152	2	0.239
264	30125 _Le	2	-0.642	1	-0.681
159	30128 _Le	2	-0.714	1	-0.902
284	30130 _Le	4	-0.133	2	0.38
160	30133 _Le	5	-0.337	4	0.093
318	30140 _Le	4	0.042	3	0.78
345	30142 _Le	2	-0.068	2	0.718
161	30150 _Le	4	0.003	2	0.551
157	30162 _Le	3	-0.56	2	-0.154
346	30168 _Le	1	-1.087	1	-0.735
283	30172 _Le	3	-0.24	2	0.208
122	30181 _Le	3	-0.371	2	0.095
121	30183 _Le	4	-0.182	3	0.127
21	30190 _Le	2	-0.681	2	-0.295
263	30196 _Le	3	0.009	2	0.754
22	30209 _Le	2	-0.786	1	-0.858
158	30220 _Le	2	-0.682	1	-1.099
224	30226 _Le	4	-0.112	2	0.357
225	30227 _Le	4	-0.139	2	0.101
223	30245 _Le	2	-0.392	2	-0.113
222	30246 _Le	6	-0.344	4	0.178
226	30247 _Le	4	-0.284	3	0.206
190	30249 _Le	5	-0.114	2	0.118
189	30250 _Le	4	-0.333	2	-0.083
193	30270 _Le	4	-0.112	2	0.439
262	30280 _Le	5	-0.045	2	0.357
280	30325 _Le	1	-0.998	1	-0.612
191	30330 _Le	3	-0.464	3	-0.011
281	30342 _Le	1	-0.964	1	-0.14
192	30394 _Le	3	-0.52	2	-0.023
282	30462 _Le	3	-0.231	3	0.48
		Mean	-0.32675	Mean	0.088875
		Standard Deviation	0.30662	Standard Deviation	0.489318

Table 11.2.2.2: Shows the calculation of means and standard deviations.

Using the means obtained in figure 11.2.2.2, the scaling constants are calculated by using equations 11.2.2.7 and 11.2.2.8 as

$$\alpha = \frac{s_{ya}}{s_{xa}} = \frac{0.30662}{0.489318} \cdot$$
$$= .626624$$

$$\beta = \bar{b}_{ya} - \alpha b_{xa}$$
$$= -0.27148 \cdot$$

The scaled item parameters of the field tested English Reading Comprehension items as displayed in figure 11.2.2.3.

Field Test Item				Item Bank Parameters of Anchor Item		Scaled		Revised	
Item No.	Label	a	b	(a_{ya})	(b_{ya})	$a_y^* = \frac{a_x}{\alpha}$	$b_y^* = \alpha b_x + \beta$	a	b
8	50060_Le	2	0.438			3	0.003	3	0.003
9	40016_Le	2	-0.024			3	-0.287	3	-0.287
10	40019_Le	3	0.12			5	-0.196	5	-0.196
11	40012_Le	1	1.4			2	0.606	2	0.606
12	40022_Le	2	-1.068			3	-0.941	3	-0.941
21	30190_Le	2	-0.295	2	-0.681	3	-0.456	3	-0.569
22	30209_Le	1	-0.858	2	-0.786	2	-0.809	2	-0.798
23	80014_Le	2	0.428			3	-0.003	3	-0.003
24	80018_Le	2	-0.024			3	-0.287	3	-0.287
25	80019_Le	2	0.12			3	-0.196	3	-0.196
31	40031_Le	2	-0.947			3	-0.865	3	-0.865
32	40049_Le	2	-0.456			3	-0.557	3	-0.557
33	40058_Le	2	-0.787			3	-0.765	3	-0.765
34	40062_Le	2	-0.911			3	-0.842	3	-0.842
35	40080_Le	3	0.012			5	-0.264	5	-0.264
41	40001_Le	3	-0.436			5	-0.545	5	-0.545
43	40003_Le	2	-1.097			3	-0.959	3	-0.959
44	40006_Le	2	-1.408			3	-1.154	3	-1.154
45	40007_Le	2	0.341			3	-0.058	3	-0.058
51	80004_Le	3	-0.055			5	-0.306	5	-0.306
52	80006_Le	1	0.361			2	-0.045	2	-0.045
53	80008_Le	3	0.169			5	-0.166	5	-0.166
54	80005_Le	2	0.045			3	-0.243	3	-0.243
55	80013_Le	3	0.045			5	-0.243	5	-0.243
60	40009_Le	2	-0.013			3	-0.280	3	-0.280
61	40035_Le	2	-0.318			3	-0.471	3	-0.471
62	40050_Le	2	-0.107			3	-0.339	3	-0.339
63	40066_Le	2	0.692			3	0.162	3	0.162
64	50062_Le	2	-0.28			3	-0.447	3	-0.447
75	40011_Le	2	0.016			3	-0.261	3	-0.261
76	40015_Le	2	-0.702			3	-0.711	3	-0.711

Field Test Item				Item Bank Parameters of Anchor Item		Scaled		Revised	
Item No.	Label	a	b	(a_{ya})	(b_{ya})	$a_y^* = \frac{a_x}{\alpha}$	$b_y^* = \alpha b_x + \beta$	a	b
77	40023_Le	3	-0.538			5	-0.609	5	-0.609
78	40025_Le	2	-1.265			3	-1.064	3	-1.064
79	40028_Le	2	-1.407			3	-1.153	3	-1.153
85	80026_Le	2	0.092			3	-0.214	3	-0.214
86	80028_Le	2	0.307			3	-0.079	3	-0.079
87	80029_Le	2	-0.28			3	-0.447	3	-0.447
88	80037_Le	3	0.281			5	-0.095	5	-0.095
89	80027_Le	2	0.673			3	0.150	3	0.150
95	40029_Le	3	-0.429			5	-0.540	5	-0.540
96	40030_Le	3	-0.64			5	-0.673	5	-0.673
97	50056_Le	3	0.109			5	-0.203	5	-0.203
98	50075_Le	2	0.646			3	0.133	3	0.133
99	40076_Le	2	-0.89			3	-0.829	3	-0.829
110	40018_Le	2	-0.592			3	-0.642	3	-0.642
111	40010_Le	2	0.459			3	0.016	3	0.016
112	40021_Le	2	-0.131			3	-0.354	3	-0.354
113	40032_Le	1	-0.278			2	-0.446	2	-0.446
114	40036_Le	2	-0.437			3	-0.545	3	-0.545
120	30078_Le	3	-0.271	4	-0.341	5	-0.441	4	-0.391
121	30183_Le	3	0.127	4	-0.182	5	-0.192	4	-0.187
122	30181_Le	2	0.095	3	-0.371	3	-0.212	3	-0.291
123	80023_Le	3	0.548			5	0.072	5	0.072
124	80021_Le	2	0.203			3	-0.144	3	-0.144
130	40046_Le	1	-1.089			2	-0.954	2	-0.954
131	40048_Le	2	0.575			3	0.089	3	0.089
133	50038_Le	3	0.01			5	-0.265	5	-0.265
134	50077_Le	2	0.609			3	0.110	3	0.110
145	40004_Le	2	-0.641			3	-0.673	3	-0.673
147	40017_Le	2	-1.565			3	-1.252	3	-1.252
148	40033_Le	2	-0.692			3	-0.705	3	-0.705
149	40056_Le	1	-0.646			2	-0.676	2	-0.676
157	30162_Le	2	-0.154	3	-0.56	3	-0.368	3	-0.464
158	30220_Le	1	-1.099	2	-0.682	2	-0.960	2	-0.821
159	30128_Le	1	-0.902	2	-0.714	2	-0.837	2	-0.775
160	30133_Le	4	0.093	5	-0.337	6	-0.213	6	-0.275
161	30150_Le	2	0.551	4	0.003	3	0.074	4	0.038
169	50001_Le	1	0.305			2	-0.080	2	-0.080
170	50078_Le	2	0.115			3	-0.199	3	-0.199
171	50079_Le	2	0.086			3	-0.218	3	-0.218
172	50052_Le	2	-0.206			3	-0.401	3	-0.401
173	50071_Le	3	0.295			5	-0.087	5	-0.087
180	40054_Le	2	-0.203			3	-0.399	3	-0.399
181	40057_Le	3	0.196			5	-0.149	5	-0.149
182	40072_Le	2	0.356			3	-0.048	3	-0.048
183	50024_Le	2	-0.147			3	-0.364	3	-0.364
189	30250_Le	2	-0.083	4	-0.333	3	-0.323	4	-0.328
190	30249_Le	2	0.118	5	-0.114	3	-0.198	4	-0.156

Field Test Item				Item Bank Parameters of Anchor Item		Scaled		Revised	
Item No.	Label	a	b	(a_{ya})	(b_{ya})	$a_y^* = \frac{a_x}{\alpha}$	$b_y^* = \alpha b_x + \beta$	a	b
191	30330_Le	3	-0.011	3	-0.464	5	-0.278	4	-0.371
192	30394_Le	2	-0.023	3	-0.52	3	-0.286	3	-0.403
193	30270_Le	2	0.439	4	-0.112	3	0.004	4	-0.054
196	50031_Le	1	-0.455			2	-0.557	2	-0.557
198	50068_Le	3	0.001			5	-0.271	5	-0.271
199	50063_Le	3	0.323			5	-0.069	5	-0.069
200	50047_Le	2	-0.824			3	-0.788	3	-0.788
212	40053_Le	2	0.34			3	-0.058	3	-0.058
213	40064_Le	1	0.327			2	-0.067	2	-0.067
214	40077_Le	1	-1.202			2	-1.025	2	-1.025
216	50030_Le	2	0.472			3	0.024	3	0.024
222	30246_Le	4	0.178	6	-0.344	6	-0.160	6	-0.252
223	30245_Le	2	-0.113	2	-0.392	3	-0.342	3	-0.367
224	30226_Le	2	0.357	4	-0.112	3	-0.048	4	-0.080
225	30227_Le	2	0.101	4	-0.139	3	-0.208	4	-0.174
226	30247_Le	3	0.206	4	-0.284	5	-0.142	4	-0.213
232	50057_Le	3	1.048			5	0.385	5	0.385
233	50040_Le	3	0.079			5	-0.222	5	-0.222
234	50041_Le	2	-0.615			3	-0.657	3	-0.657
235	50042_Le	2	0.173			3	-0.163	3	-0.163
236	50043_Le	3	0.089			5	-0.216	5	-0.216
250	40045_Le	2	0.865			3	0.271	3	0.271
251	40055_Le	2	0.28			3	-0.096	3	-0.096
252	40068_Le	3	0.184			5	-0.156	5	-0.156
253	40070_Le	3	-0.279			5	-0.446	5	-0.446
254	40071_Le	1	-0.08			2	-0.322	2	-0.322
262	30280_Le	2	0.357	5	-0.045	3	-0.048	4	-0.046
263	30196_Le	2	0.754	3	0.009	3	0.201	3	0.105
264	30125_Le	1	-0.681	2	-0.642	2	-0.698	2	-0.670
265	30098_Le	2	0.239	3	-0.152	3	-0.122	3	-0.137
266	30053_Le	2	0.608	5	-0.074	3	0.110	4	0.018
271	40073_Le	1	-0.636			2	-0.670	2	-0.670
272	50070_Le	3	0.081			5	-0.221	5	-0.221
273	50039_Le	2	-0.432			3	-0.542	3	-0.542
274	40051_Le	1	-0.774			2	-0.756	2	-0.756
275	50044_Le	2	0.105			3	-0.206	3	-0.206
276	40008_Le	1	-0.495			2	-0.582	2	-0.582
277	40034_Le	3	0.306			5	-0.080	5	-0.080
278	40038_Le	2	-0.188			3	-0.389	3	-0.389
279	40044_Le	1	0.881			2	0.281	2	0.281
280	30325_Le	1	-0.612	1	-0.998	2	-0.655	1	-0.826
281	30342_Le	1	-0.14	1	-0.964	2	-0.359	1	-0.662
282	30462_Le	3	0.48	3	-0.231	5	0.029	4	-0.101
283	30172_Le	2	0.208	3	-0.24	3	-0.141	3	-0.191
284	30130_Le	2	0.38	4	-0.133	3	-0.033	4	-0.083
291	40059_Le	2	-0.369			3	-0.503	3	-0.503
292	40075_Le	2	-0.369			3	-0.503	3	-0.503

Field Test Item				Item Bank Parameters of Anchor Item		Scaled		Revised	
Item No.	Label	a	b	(a_{ya})	(b_{ya})	$a_y^* = \frac{a_x}{\alpha}$	$b_y^* = \alpha b_x + \beta$	a	b
293	50080_Le	3	0.306			5	-0.080	5	-0.080
294	50059_Le	2	-0.004			3	-0.274	3	-0.274
305	40037_Le	2	0.229			3	-0.128	3	-0.128
306	40039_Le	2	1.507			3	0.673	3	0.673
307	40042_Le	2	0.342			3	-0.057	3	-0.057
308	50066_Le	3	0.045			5	-0.243	5	-0.243
314	17035_Le	4	0.653	3	0.018	6	0.138	5	0.078
315	21042_Le	2	0.055	2	-0.276	3	-0.237	3	-0.257
316	21055_Le	1	0.004	3	0.043	2	-0.269	2	-0.113
317	30081_Le	2	0.712	2	0.011	3	0.175	3	0.093
318	30140_Le	3	0.78	4	0.042	5	0.217	4	0.130
323	50067_Le	3	0.865			5	0.271	5	0.271
324	50037_Le	2	0.712			3	0.175	3	0.175
325	40074_Le	2	0.684			3	0.157	3	0.157
326	50050_Le	3	0.269			5	-0.103	5	-0.103
327	50054_Le	2	0.508			3	0.047	3	0.047
330	40040_Le	2	-0.343			3	-0.486	3	-0.486
331	40043_Le	2	-0.565			3	-0.626	3	-0.626
332	40047_Le	2	0.157			3	-0.173	3	-0.173
333	40060_Le	1	0.596			2	0.102	2	0.102
334	40065_Le	2	0.157			3	-0.173	3	-0.173
342	01115_Le	3	0.852	3	0.011	5	0.262	4	0.137
343	04038_Le	2	0.008	2	-0.368	3	-0.266	3	-0.317
344	30060_Le	2	0.459	5	-0.061	3	0.016	4	-0.022
345	30142_Le	2	0.718	2	-0.068	3	0.178	3	0.055
346	30168_Le	1	-0.735	1	-1.087	2	-0.732	1	-0.910
353	50026_Le	2	0.581			3	0.093	3	0.093
354	50055_Le	2	0.491			3	0.036	3	0.036
355	50045_Le	2	1.068			3	0.398	3	0.398
356	50035_Le	2	-0.257			3	-0.433	3	-0.433
357	50034_Le	3	0.388			5	-0.028	5	-0.028
358	E279_Le	1	0.524			2	0.057	2	0.057

Table 11.2.2.3: Shows the scaling of field tests items.

The field tested items from English Reading Comprehension field tests are placed on the common scale as the other items in the item bank as shown in the 'Revised' column of figure 11.2.2.2.3. These items are added to the item bank and consequently the old item bank is extended by 117 new items.

11.2.4 OPLM Method of Linking Items

OPLM uses the anchor item design to place the items from different test booklets on a common scale. The test booklets are linked by anchor items. The response data from the test booklets are combined into one data. The combination of response data from different test booklets into one response data makes it possible for OPLM to calibrate the items across booklets in single run. When the items from different booklets are calibrated simultaneously, they are calibrated on a common scale.

However, when the items from the new test booklets have to be added to the old item bank, the anchor items for the item bank and the new test booklets are used as explained in method 1 and method 2.

11.2.4.1 Method I

The response data of the new test booklets is merged with the item bank response data. Merging new response data of the test booklets with item bank response data involves defining new booklets, identifying anchor items and replacing identity numbers of the anchor items of the new response data by the identity numbers used for them in the response data of the item bank.

After successful preparation of the data, the old items in the item bank are put off except the anchor items. The parameters of the anchor items are fixed at the parameters they have in the item bank. The parameters of the new items other than the anchor items are all free. These changes are saved in a new screen file.

The new screen file is run and the resulting item parameters for the items from the new test booklets are on a common scale as that of the other items in the item bank.

11.2.4.2 Method II

The method II does not require data merging. The response data obtained from the test booklets is used for calibrating items in the booklets on a common scale as that of the items in the item bank. The parameters of the anchor items are fixed by using the parameters they have in the item bank. The parameters of the other items are set free and the screen file containing these changes is run. The resulting item parameters have common scale as that of the other items in the item bank.

After calibration, the items can be imputed into the item bank with their parameters. The item parameters generated by OPLM on a common scale as that of the other items in the item bank are shown in table 11.2.4.1.

		Item Bank		OPLM	
		Parameters of Anchor Items		Scaled	
Item No.	Label	a	b	a	b
8	50060_Le			4	0.008
9	40016_Le			3	-0.317
10	40019_Le			4	-0.24
11	40012_Le			2	0.554
12	40022_Le			3	-1.014
21	30190_Le	2	-0.681	2	-0.681
22	30209_Le	2	-0.786	2	-0.786
23	80014_Le			3	-0.014
24	80018_Le			3	-0.317
25	80019_Le			3	-0.221
31	40031_Le			3	-0.933
32	40049_Le			4	-0.476
33	40058_Le			2	-1.205
34	40062_Le			3	-0.909
35	40080_Le			4	-0.318
41	40001_Le			4	-0.662
43	40003_Le			3	-1.043
44	40006_Le			2	-1.821
45	40007_Le			2	-0.124
51	80004_Le			4	-0.377
52	80006_Le			1	-0.176
53	80008_Le			4	-0.213

		Item Bank		OPLM	
		Parameters of Anchor Items		Scaled	
Item No.	Label	a	b	a	b
54	80005_Le			3	-0.273
55	80013_Le			3	-0.376
60	40009_Le			3	-0.313
61	40035_Le			3	-0.519
62	40050_Le			3	-0.376
63	40066_Le			2	0.203
64	50062_Le			3	-0.493
75	40011_Le			3	-0.296
76	40015_Le			3	-0.779
77	40023_Le			4	-0.727
78	40025_Le			3	-1.17
79	40028_Le			3	-1.25
85	80026_Le			3	-0.244
86	80028_Le			3	-0.098
87	80029_Le			4	-0.401
88	80037_Le			5	-0.107
89	80027_Le			2	0.177
95	40029_Le			4	-0.646
96	40030_Le			4	-0.803
97	50056_Le			4	-0.248
98	50075_Le			3	0.133
99	40076_Le			4	-0.727
110	40018_Le			3	-0.702
111	40010_Le			2	-0.036
112	40021_Le			3	-0.406
113	40032_Le			2	-0.356
114	40036_Le			3	-0.599
120	30078_Le	4	-0.341	4	-0.341
121	30183_Le	4	-0.182	4	-0.182
122	30181_Le	3	-0.371	3	-0.371
123	80023_Le			4	0.058
124	80021_Le			3	-0.172
130	40046_Le			2	-0.772
131	40048_Le			3	0.076
133	50038_Le			4	-0.331
134	50077_Le			3	0.098
145	40004_Le			2	-1.173
147	40017_Le			3	-1.383
148	40033_Le			3	-0.801
149	40056_Le			2	-0.561
157	30162_Le	3	-0.56	3	-0.56
158	30220_Le	2	-0.682	2	-0.682
159	30128_Le	2	-0.714	2	-0.714
160	30133_Le	5	-0.337	5	-0.337
161	30150_Le	4	0.003	4	0.003
169	50001_Le			2	-0.076
170	50078_Le			2	-0.447
171	50079_Le			3	-0.283
172	50052_Le			2	-0.751

		Item Bank		OPLM	
		Parameters of Anchor Items		Scaled	
Item No.	Label	a	b	a	b
173	50071_Le			5	-0.128
180	40054_Le			4	-0.388
181	40057_Le			5	-0.229
182	40072_Le			3	-0.153
183	50024_Le			3	-0.483
189	30250_Le	4	-0.333	4	-0.333
190	30249_Le	5	-0.114	5	-0.114
191	30330_Le	3	-0.464	3	-0.464
192	30394_Le	3	-0.52	3	-0.52
193	30270_Le	4	-0.112	4	-0.112
196	50031_Le			2	-0.492
198	50068_Le			5	-0.345
199	50063_Le			4	-0.202
200	50047_Le			3	-0.929
212	40053_Le			3	-0.147
213	40064_Le			2	-0.065
214	40077_Le			2	-0.847
216	50030_Le			3	-0.058
222	30246_Le	6	-0.344	6	-0.344
223	30245_Le	2	-0.392	2	-0.392
224	30226_Le	4	-0.112	4	-0.112
225	30227_Le	4	-0.139	4	-0.139
226	30247_Le	4	-0.284	4	-0.284
232	50057_Le			4	0.312
233	50040_Le			4	-0.367
234	50041_Le			3	-0.781
235	50042_Le			3	-0.258
236	50043_Le			4	-0.36
250	40045_Le			2	0.081
251	40055_Le			2	-0.461
252	40068_Le			5	-0.275
253	40070_Le			4	-0.687
254	40071_Le			2	-0.323
262	30280_Le	5	-0.045	5	-0.045
263	30196_Le	3	0.009	3	0.009
264	30125_Le	2	-0.642	2	-0.642
265	30098_Le	3	-0.152	3	-0.152
266	30053_Le	5	-0.074	5	-0.074
271	40073_Le			2	-0.603
272	50070_Le			4	-0.426
273	50039_Le			3	-0.728
274	40051_Le			2	-0.672
275	50044_Le			3	-0.375
276	40008_Le			2	-0.452
277	40034_Le			5	-0.149
278	40038_Le			3	-0.509
279	40044_Le			2	0.254
280	30325_Le	1	-0.998	1	-0.998
281	30342_Le	1	-0.964	1	-0.946

		Item Bank		OPLM	
		Parameters of Anchor Items		Scaled	
Item No.	Label	a	b	a	b
282	30462_Le	3	-0.231	3	-0.231
283	30172_Le	3	-0.24	3	-0.24
284	30130_Le	4	-0.133	4	-0.133
291	40059_Le			3	-0.629
292	40075_Le			3	-0.629
293	50080_Le			4	-0.225
294	50059_Le			3	-0.387
305	40037_Le			3	-0.228
306	40039_Le			3	0.641
307	40042_Le			3	-0.153
308	50066_Le			4	-0.408
314	17035_Le	3	0.018	3	0.018
315	21042_Le	2	-0.276	2	-0.276
316	21055_Le	3	0.043	3	0.043
317	30081_Le	2	0.011	2	0.011
318	30140_Le	4	0.042	4	0.042
323	50067_Le			5	0.202
324	50037_Le			3	0.096
325	40074_Le			3	0.077
326	50050_Le			4	-0.243
327	50054_Le			3	-0.042
330	40040_Le			3	-0.659
331	40043_Le			3	-0.808
332	40047_Le			3	-0.325
333	40060_Le			2	0.091
334	40065_Le			3	-0.325
342	01115_Le	3	0.011	3	0.011
343	04038_Le	2	-0.368	2	-0.368
344	30060_Le	5	-0.061	5	-0.061
345	30142_Le	2	-0.068	2	-0.068
346	30168_Le	1	-1.087	1	-1.087
353	50026_Le			2	-0.2
354	50055_Le			2	-0.285
355	50045_Le			3	0.291
356	50035_Le			3	-0.602
357	50034_Le			4	-0.213
358	E279_Le			2	0.03

Table 11.2.4.1: Shows the item parameters after linking by using OPLM.

11.2.5 Comparison of Mean and Sigma and OPLM methods of Linking Items

To compare the parameters of the items linked to the items in the item bank by using Mean and Sigma method and OPLM method, an assumption is made that the item parameters obtained by different methods should have linear relationship as they are already on the common scale. To test this assumption, the item discrimination parameters from different methods are compared by using linear plots. Similarly, the item difficulty parameters from different methods are compared by using linear plots as well.

Figure 11.2.5.1 shows the linear plot of **ams** (item discrimination parameter from mean and sigma method) against **aoplm** (item discrimination parameter from OPLM method) with 95% percent confidence interval.

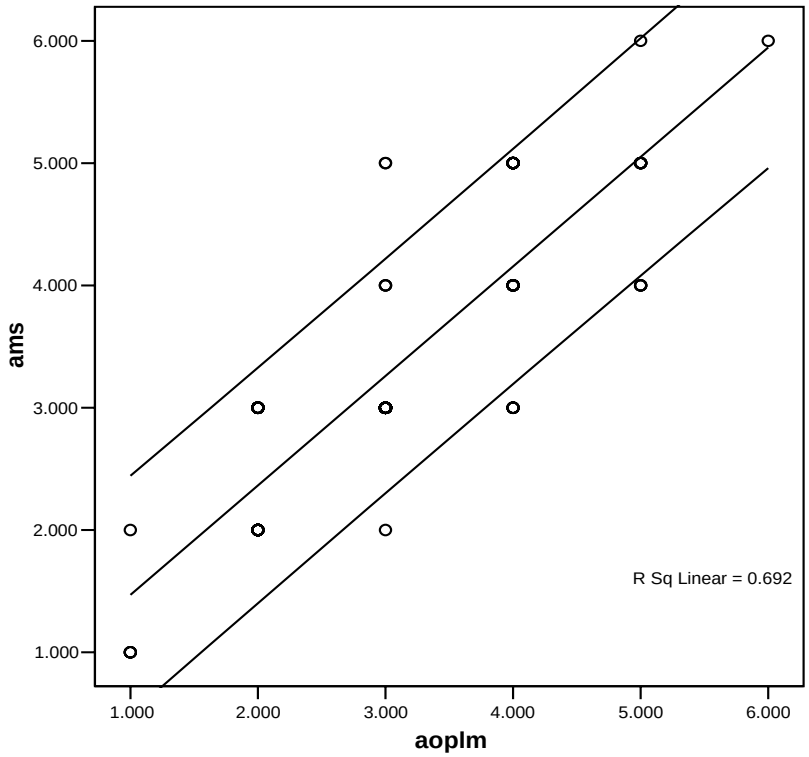

Figure 11.2.5.1: Shows the plot of **ams** against **aoplm**.

Figure 11.2.5.2 shows the linear plot of **bms** (item difficulty parameter from mean and sigma method) and **boplm** (item difficulty parameter from OPLM method).

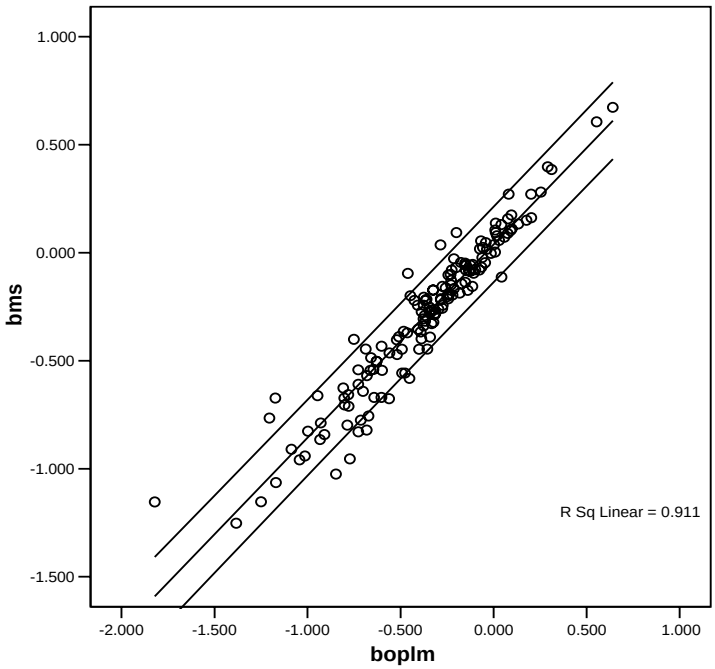

Figure 11.2.5.2: Shows the linear plot of **bms** and **boplm**.

From figure 11.2.5.1, it can be inferred that the item discriminations from mean and sigma method and from OPLM show a fair linear relationship. A difference is expected because the mean and sigma method allows for sample fluctuation where as OPLM method controls sample fluctuation by fixing the parameters of the anchor items.

Figure 11.2.5.2 shows a good relationship between item difficulty parameters from mean and sigma method and OPLM method. However, they vary to certain extent and this is due to the effect of sampling fluctuation in mean and sigma method contrary to its absence in OPLM method.

To assess the superiority of one method over the other, a third reference is desirable. However, this is out of the scope of the thesis.

11.2.6 Summary

Item banks are the files of various suitable test items that are coded by subject area, instructional level, instructional objective measured and various pertinent item characteristics like item difficulty and item discriminating power.

Item bank can assist a test developer to (a) do item entry and storage (b) do item retrieval for reviewing items, formatting test forms, and editing and updating items and (c) maintain item history.

Anchor item design uses common items to link different test forms or test booklets.

Mean and Sigma Method of determing scaling constants uses the assumption of linear relationships between the parameters of anchor items in different test forms.

The parameters of the anchor items become different due to sample fluctuation after scaling. To adjust the differences, the parameters of the anchor items after scaling and their original parameters are averaged and the values obtained will replace the old parameters.

11.2.7 References

Gronlund, E.N. (1998, 6th ed.,). *Assessment of Student Achievement*: Alyn and Bacon

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

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

Ward,W. Annie (1994). An NCME Instructional Module on Guidelines for the Development of Item Banks. In *Educational Measurement, ITEMS*: National Council on Measurement in Education, Princeton, NJ.

Rudner, Lawrence (1998). Item banking. *Practical Assessment, Research & Evaluation*, 6(4). Retrieved July 6, 2006 from http://PAREonline.net/getvn.asp?v=6&n=4

Chapter 12

Discussions and Future Developments

12.0.0 Introduction

In the previous 11 chapters, I have shown how CTT and IRT are applied in test constructions, DIF studies and item banking. The thesis extended its scope into the Dutch Student Monitoring Systems which use test statistics as the core of information for making decisions related to the students of the Dutch Primary Education Schools and Secondary Education Schools respectively.

In this chapter, some anecdotes from some of the previous chapters will be presented for discussions followed by brief outlines of the likelihood of using IRT at the Bhutan Board of Examinations, Ministry of Education, Royal Government of Bhutan, Bhutan.

12.1.0 Discussions

The use of IRT as a means for generating valid and precise information about students' learning competencies is not only promising, but also being widely practised in the community of testing centers across the globe. As a matter of fact, IRT is a necessary tool which has to be available at any testing centers should they want to make valid and reliable large scale test instruments.

The fundamental structure of the Dutch Education System and two important tools known as Student Monitoring Systems to regulate its quality at Primary Education and Secondary Education are described in chapter 1. The longitudinal monitoring design element is the building block of the Student Monitoring Systems. The students' performances in different tests in the same subject areas in different times provide the policy makers with valuable information about the progress in students' learning. The variables involved in UIBTERV are assumed to be sufficient to take into account all possible reasons which decelerate and accelerate students' learning progress. The identified reasons have the potential to assist teachers and policy makers to reconsider their roles and positions as defined with reference to the making of Dutch Education System. It would be interesting to make an impact study about the changes brought about in the Dutch Education System by the Student Monitoring Systems.

IRT is known to be replacing CTT, but there are people in the community of measurement specialists who favor the combined use of IRT and CTT. OPLM performs both CTT and IRT analyses. The CTT statistics aid in the study of model fit when an IRT model is used for studying the response data.

The versatility of OPLM to take different forms of IRT models like Rasch model, two parameter logistic model, partial credit model and generalized partial credit model makes it suitable for use in studying a wide range of response data. The simplicity of conducting DIF studies is yet another important feature that OPLM has. It is known that the DIF statistics generated by OPLM are similar to the DIF statistics generated by Restricted Factor Analysis, Mantel-Haenszel Method and Likelihood Ratio Method. The capability of OPLM to link items to an item bank extends its function in building as well as extending an item bank.

12.2.0 Future Developments

Despite its popularity among the developed and well equipped testing centers, IRT is still a nascent subject in many testing centers. The reasons are not hard to understand. To use IRT both effectively and meaningfully, expertise to (a) define the purpose behind the use of IRT, (b) define latent space, (c)

confirm goodness of fit between the model and data, (d) decide the direction of adjustment between data and model, (e) confirm the sufficiency of sample size, (f) run computer programs, (g) decide on the appropriateness of the estimation procedures, (h) interpret test scores and (i) assess the validity of the information provided by the IRT is required. This is an enormous demand.

The Bhutan Board of Examinations, Ministry of Education, Royal Government of Bhutan, Bhutan has a keen interest and strong policy goals to (a) provide fair national examinations to the students and (b) present valid and reliable information about the health of the education system to the Ministry of Education, Royal Government of Bhutan. These two broad goals are largely being achieved by studying the performances of the students in various national examinations. The use of IRT as the tool to explore the wealth of information contained in the students' performances in various national examinations is not only desirable but also a necessity. As a matter of fact, Bhutan Board of Examinations has been developing its capacity by purchasing consultancy services and short term trainings from abroad as partial fulfillment of its long term policy of building and enhancing staff capacity.

The Student Monitoring Systems constructed by using IRT may become a strong focus among the Bhutanese educationists. A kind of system similar to Student Monitoring Systems known as National Educational Assessment exists in Bhutan. There are several similarities between the two systems. In this regard, Bhutan Board of Examinations will be looking forward to working closely with CITO.

12.3.0 Summary

IRT is a necessary tool which has to be available at any testing centers should they want to make valid and reliable large scale test instruments.

The longitudinal monitoring design element is the building block of the Dutch Student Monitoring Systems.

It would be interesting to make an impact study about the changes brought about in the Dutch Education System by the Student Monitoring Systems.

Despite its popularity among the developed and well equipped testing centers, IRT is still a nascent subject in many testing centers.

A kind of system similar to Student Monitoring Systems known as National Educational Assessment exists in Bhutan.

Appendix I: Global Norm for UIBTERV English Reading Comprehension Test

Conditional Distributions				Score Distributions						
Score	Theta	Mean	St.Dev.	ALL	Boys	Girls	Unknown	BB+/BB	KB/GL	HA/VW
0	-4.257	-3.946	0.408	0	0	0	0	0	0	0
1	-3.711	-3.711	0.513	0	0	0	0	0	0	0
2	-3.085	-3.245	0.582	0	0	0	0	0	0	0
3	-2.736	-2.887	0.546	0	0	0	0	0	0	0
4	-2.495	-2.615	0.481	0	0	0	0	0	0	0
5	-2.313	-2.406	0.417	0	0	0	0	0	0	0
6	-2.166	-2.239	0.364	0	0	0	0	0	0	0
7	-2.043	-2.101	0.323	0	0	0	0	0	0	0
8	-1.937	-1.985	0.291	0	0	0	0	0	0	0
9	-1.843	-1.884	0.266	0	0	0	0	0	0	0
10	-1.76	-1.795	0.246	0	0	0	0	0	0	0
11	-1.684	-1.715	0.23	0	0	0	0	0	0	0
12	-1.615	-1.643	0.217	0	0	0	0	0	0	0
13	-1.552	-1.576	0.206	0	0	0	0	0	0	0
14	-1.492	-1.515	0.196	0	0	0	0	0	0	0
15	-1.437	-1.458	0.188	0	0	0	0	0	0	0
16	-1.386	-1.404	0.181	0	0	0	0	0	0	0
17	-1.337	-1.354	0.174	0	0	0	0	0	0	0
18	-1.291	-1.307	0.168	0	0	0	0	0	0	0
19	-1.247	-1.262	0.163	0	0	0	0	0	0	0
20	-1.205	-1.219	0.158	0	0	0	0	0	0	0
21	-1.165	-1.178	0.154	0	0	0	0	0	0	0
22	-1.127	-1.139	0.15	0	0	0	0	0	0	0
23	-1.091	-1.102	0.146	0	0	0	0	0	0	0
24	-1.055	-1.066	0.142	0	0	0	0	0	0	0
25	-1.022	-1.032	0.139	0	0	0	0	0	0	0
26	-0.989	-0.999	0.136	0	0	0	0	0	0	0
27	-0.958	-0.967	0.133	0	0	0	0	0	0	0
28	-0.927	-0.936	0.131	0	0	0	0	0	0	0
29	-0.898	-0.906	0.128	0	0	0	0	0	0	0
30	-0.869	-0.877	0.126	0	0	0	0	0	0	0
31	-0.841	-0.849	0.124	0	0	0	0	0	0	0
32	-0.814	-0.821	0.122	0	0	0	0	0	0	0
33	-0.788	-0.795	0.12	0	0	0	0	0	0	0
34	-0.762	-0.769	0.118	0	0	0	0	0	0	0
35	-0.737	-0.744	0.116	0	0	0	0	0	0	0
36	-0.713	-0.719	0.114	0	0	0	0	0	0	0
37	-0.689	-0.695	0.113	0	0	0	0	0	0	0
38	-0.665	-0.671	0.111	0	0	0	0	1	0	0
39	-0.643	-0.648	0.11	0	0	0	0	1	0	0
40	-0.62	-0.625	0.109	0	0	0	0	1	0	0
41	-0.598	-0.603	0.107	0	0	0	0	1	0	0
42	-0.577	-0.582	0.106	0	0	0	0	1	0	0

Conditional Distributions				Score Distributions						
Score	Theta	Mean	St.Dev.	ALL	Boys	Girls	Unknown	BB+/BB	KB/GL	HA/VW
43	-0.556	-0.56	0.105	0	1	0	0	1	0	0
44	-0.535	-0.539	0.104	1	1	0	0	1	0	0
45	-0.514	-0.519	0.103	1	1	0	0	1	0	0
46	-0.494	-0.499	0.102	1	1	0	0	1	0	0
47	-0.475	-0.479	0.101	1	1	1	0	2	0	0
48	-0.455	-0.459	0.1	1	1	1	1	2	1	0
49	-0.436	-0.44	0.099	1	1	1	1	2	1	0
50	-0.417	-0.421	0.098	1	1	1	1	2	1	0
51	-0.399	-0.402	0.097	1	1	1	1	2	1	0
52	-0.38	-0.384	0.096	1	1	1	1	2	1	0
53	-0.362	-0.365	0.095	1	1	1	1	3	1	0
54	-0.344	-0.347	0.095	1	1	1	1	3	1	0
55	-0.327	-0.33	0.094	1	2	1	1	3	1	0
56	-0.309	-0.312	0.093	2	2	1	1	3	1	0
57	-0.292	-0.295	0.093	2	2	1	1	4	1	0
58	-0.275	-0.277	0.092	2	2	2	1	4	1	0
59	-0.258	-0.26	0.092	2	2	2	1	4	2	0
60	-0.241	-0.243	0.091	2	2	2	2	5	2	0
61	-0.225	-0.227	0.091	2	3	2	2	5	2	0
62	-0.208	-0.21	0.09	3	3	2	2	5	2	0
63	-0.192	-0.194	0.09	3	3	2	2	6	2	0
64	-0.176	-0.178	0.089	3	3	2	2	6	2	0
65	-0.16	-0.161	0.089	3	3	3	2	7	3	0
66	-0.144	-0.145	0.088	3	3	3	2	7	3	0
67	-0.128	-0.129	0.088	4	4	3	3	8	3	0
68	-0.112	-0.114	0.088	4	4	3	3	8	3	0
69	-0.097	-0.098	0.087	4	4	3	3	9	3	1
70	-0.081	-0.082	0.087	4	4	4	3	9	4	1
71	-0.066	-0.067	0.087	5	5	4	3	10	4	1
72	-0.05	-0.051	0.086	5	5	4	4	10	4	1
73	-0.035	-0.036	0.086	5	5	5	4	11	4	1
74	-0.02	-0.021	0.086	5	6	5	4	12	5	1
75	-0.005	-0.005	0.086	6	6	5	4	12	5	1
76	0.011	0.01	0.086	6	6	5	5	13	5	1
77	0.026	0.025	0.085	6	7	6	5	13	6	1
78	0.041	0.04	0.085	7	7	6	5	14	6	1
79	0.056	0.055	0.085	7	7	7	6	15	7	1
80	0.071	0.07	0.085	8	8	7	6	16	7	1
81	0.086	0.085	0.085	8	8	7	6	16	7	1
82	0.101	0.1	0.085	8	8	8	7	17	8	2
83	0.115	0.115	0.085	9	9	8	7	18	8	2
84	0.13	0.13	0.085	9	9	9	7	19	9	2
85	0.145	0.145	0.085	10	10	9	8	20	9	2
86	0.16	0.16	0.085	10	10	10	8	21	10	2
87	0.175	0.176	0.085	11	11	10	9	22	11	2
88	0.19	0.191	0.085	11	11	11	9	23	11	2
89	0.205	0.206	0.085	12	12	11	10	24	12	3
90	0.22	0.221	0.085	13	12	12	10	25	12	3
91	0.235	0.236	0.085	13	13	13	11	26	13	3
92	0.25	0.251	0.085	14	13	13	11	27	14	3

Conditional Distributions				Score Distributions						
Score	Theta	Mean	St.Dev.	ALL	Boys	Girls	Unknown	BB+/BB	KB/GL	HA/VW
93	0.266	0.267	0.086	14	14	14	12	28	15	3
94	0.281	0.282	0.086	15	15	15	12	29	15	4
95	0.296	0.297	0.086	16	15	15	13	30	16	4
96	0.312	0.313	0.086	16	16	16	14	31	17	4
97	0.327	0.328	0.087	17	17	17	14	32	18	5
98	0.343	0.344	0.087	18	17	18	15	34	19	5
99	0.358	0.36	0.087	19	18	18	16	35	20	5
100	0.374	0.376	0.087	20	19	19	16	36	21	5
101	0.39	0.392	0.088	20	20	20	17	37	22	6
102	0.406	0.408	0.088	21	21	21	18	39	23	6
103	0.422	0.424	0.089	22	21	22	19	40	24	7
104	0.438	0.44	0.089	23	22	23	20	41	25	7
105	0.455	0.457	0.09	24	23	24	21	43	26	8
106	0.471	0.474	0.09	25	24	25	22	44	27	8
107	0.488	0.49	0.091	26	25	26	22	46	28	9
108	0.505	0.507	0.091	27	26	27	23	47	29	9
109	0.522	0.525	0.092	28	27	29	24	49	31	10
110	0.539	0.542	0.092	29	28	30	26	50	32	10
111	0.557	0.56	0.093	31	29	31	27	52	33	11
112	0.574	0.578	0.094	32	30	32	28	53	35	12
113	0.592	0.596	0.095	33	32	34	29	55	36	12
114	0.611	0.614	0.095	34	33	35	30	56	38	13
115	0.629	0.633	0.096	36	34	36	31	58	39	14
116	0.648	0.652	0.097	37	35	38	33	59	41	15
117	0.667	0.671	0.098	38	36	39	34	61	42	16
118	0.686	0.691	0.099	40	38	41	35	62	44	17
119	0.706	0.711	0.1	41	39	42	37	64	46	18
120	0.726	0.731	0.101	43	41	44	38	66	47	19
121	0.747	0.752	0.102	44	42	46	40	67	49	20
122	0.768	0.773	0.104	46	44	47	41	69	51	21
123	0.789	0.795	0.105	47	45	49	43	70	53	23
124	0.811	0.817	0.106	49	47	51	45	72	54	24
125	0.833	0.839	0.108	51	48	53	46	74	56	25
126	0.856	0.863	0.109	52	50	54	48	75	58	27
127	0.88	0.886	0.111	54	52	56	50	77	60	29
128	0.904	0.911	0.113	56	53	58	52	78	62	30
129	0.929	0.936	0.115	58	55	60	54	80	64	32
130	0.954	0.962	0.117	60	57	62	56	81	66	34
131	0.98	0.989	0.119	62	59	64	58	83	68	36
132	1.008	1.017	0.122	64	61	66	60	84	70	38
133	1.036	1.045	0.124	66	63	68	62	86	72	40
134	1.065	1.075	0.127	68	65	70	64	87	74	43
135	1.095	1.106	0.13	70	67	72	66	88	76	45
136	1.127	1.139	0.134	72	69	74	68	89	78	48
137	1.16	1.172	0.137	74	71	76	70	91	80	50
138	1.194	1.208	0.142	76	73	78	73	92	82	53
139	1.231	1.245	0.146	78	75	80	75	93	84	56
140	1.269	1.285	0.151	80	77	82	77	94	86	59
141	1.309	1.327	0.157	82	79	84	79	95	88	62
142	1.352	1.371	0.163	84	81	86	82	96	89	65

Conditional Distributions				Score Distributions						
Score	Theta	Mean	St.Dev.	ALL	Boys	Girls	Unknown	BB+/BB	KB/GL	HA/VW
143	1.398	1.419	0.171	86	83	88	84	96	91	68
144	1.447	1.47	0.179	88	86	90	86	97	92	72
145	1.501	1.526	0.189	90	88	92	88	98	94	75
146	1.558	1.587	0.201	91	89	93	90	98	95	79
147	1.622	1.655	0.216	93	91	95	92	99	96	82
148	1.693	1.731	0.234	94	93	96	94	99	97	85
149	1.772	1.817	0.257	96	95	97	95	99	98	88
150	1.863	1.917	0.286	97	96	98	96	100	99	91
151	1.969	2.036	0.324	98	97	99	98	100	99	94
152	2.096	2.182	0.371	99	98	99	98	100	100	96
153	2.255	2.366	0.428	99	99	100	99	100	100	98
154	2.467	2.605	0.485	100	100	100	100	100	100	99
155	2.777	2.921	0.514	100	100	100	100	100	100	100
156	3.347	3.329	0.446	100	100	100	100	100	100	100
157	3.792	3.512	0.36	100	100	100	100	100	100	100

Table 9.2.0.1: Shows item information and item p-values.