2008

Computerized adaptive testing in mathematics for primary schools in Thailand

Chaowprapha Chuesathuchon
Edith Cowan University

Edith Cowan University

Copyright Warning

COMPUTERIZED ADAPTIVE TESTING IN MATHEMATICS FOR PRIMARY SCHOOLS IN THAILAND

By

Chaowprapha Chuesathuchon
Grad.Dip.Sci. (Interdisciplinary Studies)
M.Ed. (Educational Test & Measurement)
B.Ed. (Mathematics)

This thesis is presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Education and Arts
Edith Cowan University

Date of submission: September, 2008

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

ABSTRACT

A new system-wide educational change has been introduced in Thailand requiring, amongst other things, that technologies for education be utilized in all levels of education. This study focuses on Mathematics Computerized Adaptive Testing (CAT) designed to provide Thai primary school teachers with an innovative method of assessment, one of the effective tools of new technology to be used in Thai primary schools. This study aims to: (1) construct multiple-choice test items for a Mathematics course on the topic of Equations for Year 6 (Prathom Suksa 6) students in Thailand; (2) calibrate the test items using a Rasch Measurement Model, investigate the validity and reliability of the test data, and set up the items in an item bank for use with Computerized Adaptive Testing; (3) create a computer program for Computerized Adaptive Testing, test the program and modify the program as appropriate; (4) construct and develop an attitudinal questionnaire about the Mathematics Computerized Adaptive Testing; (5) investigate the Mathematics abilities and attitudes to the Mathematics Computerized Adaptive Testing of Thailand Year 6 students; (6) compare the test length, testing time, and mathematics competency for different stopping criteria; and (7) compare the test length and testing time among differences in mathematics competency of the examinees.

The study was conducted in three data parts (creating an item bank, Computerized Adaptive Testing and attitude to CAT). In the first part, 290 multiple-choice test items on mathematical equations were created for an item bank for use in part two. They consisted of nine aspects: (1) identifying an equation; (2) identifying the true equation; (3) identifying equations with an unknown; (4) finding the value of an unknown that satisfies the equation; (5) identifying the method to solve the equation; (6) finding the solutions to equations; (7) finding a solution of an equation which related the given condition; (8) selecting an equation converted from a verbal problem or a verbal problem related to an equation; and (9) solving the problem. A total of 290 items of seven papers with 50 items each, and each paper contained 40 different items and 10 common items administered to 3,062 students of Year 6 (Prathom Suksa 6). There were 409, 413, 412, 400, 410, 408, and 610 students taking part in the 1st to the 7th tests respectively. The data were analysed with the Rasch Unidimensional Measurement Model (RUMM 2010) computer program. Ninety-eight test items fitted the measurement model and were installed in the item bank.

In part two, a computer program for Mathematics Computerized Adaptive Testing was created, tested, and modified after trialling. A controlled experiment involving 400 Prathom Suksa 6 students from two primary schools in Ubon Ratchathani province, Thailand, was implemented. The gender-ability mix of students from each school were randomly assigned to four subgroups. Each group contained 100 members whose different mathematical competencies were mixed: 30 students in high ability group, 40 in medium, and 30 in low competency level. Four stopping criteria techniques were used for simple random selection of the students for each group with the SPSS computer program. A one-way ANOVA was used to examine differences in test length and testing times among the different groups relating to stopping criteria and mathematics competencies, and also to examine differences in mathematical

competencies among the different groups of stopping criteria. Results indicated that: (1) the item bank of equations for the Prathom Suksa 6 students contained 98 items which fitted the measurement model and consisted of nine aspects, ordered from very easy (-1.27 logits) to very hard (+1.57 logits); (2) test lengths, testing times, and mathematical competencies were significantly different at p=0.05 among four groups of stopping criteria ($SEE \leq 0.20$; $SEE \leq 0.30$, $SEE \leq 0.40$ and $SEE_m - SEE_{m-1} \leq 0.005$); (3) test lengths and testing times were significantly different at p=0.05 among the three groups of mathematical competencies (low, moderately high, and high); and (4) there were 72.25 %, 16.75%, and 8% of the Prathom Suksa 6 students having a moderately high, low, and high mathematics achievement respectively.

In part three, the RUMM 2010 computer program was used to create a linear scale of Student Attitude towards Computerized Adaptive Testing. Attitude was conceptualised from five aspects: (1) Like and Interest in CAT; (2) Confidence with and Use of CAT; (3) CAT as Modern and Useful; (4) CAT is Reliable; and (5) CAT Recommendations. Data were collected from 400 Prathom Suksa 6 students and an interval scale was created with 30 items (27 items fitted the measurement model with probability p>0.04) and there was acceptable overall fit (the item-trait chi-square = 165.4, df =150, p=0.18). Results indicated that students had a very positive attitude towards computerized adaptive testing for mathematics in primary school. The three easiest items were: I am ready to apply the knowledge from CAT, CAT gives reliable results and CAT is very useful. The three hardest items were: I took CAT with confidence, I believe that I can do CAT well, and CAT saves money.

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher education.

(ii) contain any material previously published or written by another person except where due reference is made in the text; or

(iii) contain any defamatory material.

I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Signature:

Date: September 29, 2008

ACKNOWLEDGEMENTS

I would like to acknowledge the enormous support and guidance offered by my supervisor, Professor Dr Russell F. Waugh. His suggestions, insight, and constructive criticisms have provided the necessary inspiration to enable me to continue on this thesis.

My sincere thanks also go to the directors of the target schools and students who willingly participated in my research either at the pilot or data collection phase.

I would also like to express my thanks to former President, Assistant Professor Somchai Wongkasem, the present president, Assistant Professor Kasem Boonrom, colleagues at Ubon Ratchathani Rajabhat University and friends whose assisted with lively debates were regular encouragement.

I greatly appreciate Assistant Professor Dr Chompunoot Morachart, Dr Chayada Danuwong, Assistant Professor Dr Kittima Cheungsuwadi, Assistant Professor Dr Nalinee Thongprasert, Associate Professor Prayong Thitithananon, Aajarn Udomdej Tharahom, and Khun Jaturabhut Imwut who distributed their expertise.

Finally, I wish to sincerely thank my husband, Associate Professor Dr Chuanchai, and children, Yuwat, Padcha, and Nisakorn for their love, understanding, patience, and practical assistance during my years of study. Without their unconditional support this thesis would have never been completed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter introduces the reader to the Thai Educational System and the background of assessment. The problems related to assessment in Thailand are described. The significance of the study, the purpose of the study, research questions and definition of terms, are also presented.

The Thai Educational System

The Structure of the Education System

The basic structure of Thai education is twelve years basic education guaranteed by the Constitution of 1997 and provided free. Of this, nine years are compulsory. The National Education Act of 1999 (Office of the National Education Commission, 1999) was introduced to implement the constitutional right of Thai citizens to twelve years of free schooling. This objective is to be achieved through formal, non-formal and informal education (Office of the National Education Commission, 1999, p.7).What is of concern to educators in Thailand is the formal school system, and within it, the role of mathematics and assessment of student achievement.

A curriculum framework has been developed for implementation in primary grades 1-3 (Prathom Suksa 1-3), primary grades 4-6 (Prathom Suksa 4-6), secondary grades 7-9 (Mathayom Suksa 1-3), and, finally, secondary grades 10-12 (Mathayom Suksa 4-6), in line with Sections 27-28 of the Education Act (Office of the National Education Commission, 1999, p.12). The new curriculum was introduced in the academic year 2002 to be implemented in all grades by 2004, a review being scheduled for 2005. This curriculum emphasizes Mathematics as one of the most important subjects, taught from grade 1 in all grades. The Act itself emphasizes the importance in Section 23, point (4), where it states that education "shall give emphases to knowledge and skills in mathematics and languages, with emphasis on proper use of the Thai language" (Office of the National Education Commission, 1999, p.10). Mathematics is emphasized together with language knowledge, as properly grounded in the Thai

language. According to chapter 9, Technologies for Education, of the National Education Act of B.E.2542 (1999) in Thailand, the state shall promote and support the production and development of technologies for education, educate both producers and users for technology literacy, promote research and development, and the production and refinement of technologies for education, and learners shall have the right to develop their capabilities for utilization of technology for education (Office of the National Education Commission, 1999). These are relevant to the design of this research, as will be seen later.

The Background to the Assessment

Assessment is one of the most important factors of the Thai educational system. It is an integral part of the learning process. The major purpose of assessment is to improve learning (The Institute for the Promotion of Teaching Science and Technology, 2004). Assessment of student achievement is an important factor in schooling. Assessment is a major concern in education as test scores are an index of academic achievement (Richichi, 1996). The results of assessment can predict whether an education product is successful or not. The reliability of the outcome using assessment tools is important to students and teachers (Wiboonsri, 2005). There are many kinds of educational assessment tools. They are observation, interview, questionnaire, test and so on. The most popular one used by teachers is an achievement test. Most teachers try to increase efficiency in measuring achievement (Karnjanawasri, 2005), but many, if not most, lack expertise in measurement.

Three major problems relating to the assessment in Thailand come from teachers, testing, and the students. For example, many Thai teachers lack knowledge about how to construct good test items, and they have difficulty in getting help. They can not turn to item test banks to help them because there are none available for the teachers (The Institute for the Promotion of Teaching Science and Technology, 2004). They can not choose the tools to suit the learning objectives and the content. They do not know how to calibrate test items, and they do not realize the significance of constructing and calibrating test items. In regard to testing problems, appropriate test items and item banks are not available in many subjects for primary schools in Thailand. The test items which are used in schools are often of poor quality. Testing arrangements are often inappropriate, and test items sometimes do not indicate the actual abilities of the students with sufficient accuracy. Students in Thai primary

schools do not often prepare themselves for testing (The Institute for the Promotion of Teaching Science and Technology, 2004).

In traditional testing, all students in a class have to do the same test in the same time, irrespective of individual differences in students' abilities. Using the same test usually causes some problems, as some items are too easy, or too hard, for someone whose abilities are quite low, or high. As a result, students sometimes guess the answers for testing, spend too much time on some items, are bored, and are careless in doing test items. Such problems cause errors in measurement (Maneelek, 1997).

The Thai educational system is not satisfactorily successful, partly due to the tests which are of the knowledge and memory type, and the development of measuring and evaluating tools, which correspond to the development of shills and abilities of learners, are now needed. For the educational development to be successful, it is necessary to improve the assessment tools.

Nowadays, it is generally assumed that Thai teachers are not prepared to create the tools to evaluate learning achievement (The Institute for the Promotion of Teaching Science and Technology, 2004). Therefore, most tests in Thailand are objective tests focusing on knowledge and memory. Thus, there is a need to develop the tools to evaluate learning achievement in Thai schools and effective tools can lead to improvement in the Thai educational system (The Institute for the Promotion of Teaching Science and Technology, 2004).

The present study on mathematics computerized adaptive testing was designed to provide Thai primary school teachers with an innovative method of assessment which is one affective tool of new technologies for use in primary schools in Thailand.

Lord and Novick (1968, p.359) suggested that a test used for assessment should be suitable for examinees' abilities. Therefore, an adaptive test, in which items, selected from a large bank of items, close to the ability of a particular examinee (Stocking & Swanson, 1998), is one alternative instrument to solve the assessment problems.

Davey and Parshall (cited in Stocking & Swanson, 1998, p.271) noted that:

> Adaptive testing has three goals: (1) to maximize test efficiency by selecting the most appropriate items for an examinee, (2) to ensure that tests measure the same traits for each examinee by controlling the non-statistical nature of test

items, and (3) to protect the security of the item bank by controlling the rates at which items are administered (p.271).

At present, almost all adaptive testing designs assume that items will be selected and administered by a computer and the test items calibrated by using Item Response Theory (IRT). Technology and assessment are connected through Computerized Adaptive Testing (CAT) (Lord, Wainer, Green, Mislevy, & Thissen cited in Stocking & Swanson, 1998, p.271). CAT is a test that tailors the assessment process by choosing test items which are close to a person's ability level (Shermis, Stemmer, & Webb, 1996). This means that questions which are too easy or too hard for an individual are avoided, and the test ends as soon as an a priori confidence level is reached. Computerized Adaptive Tests tend to be shorter than conventional fixed-length tests without loss of reliability (Kyungsu, 1996).

Meijer and Nering (1999, p.18) noted that:

> The objective of computerized adaptive testing is to construct an optimal test for each examinee. To achieve this, an examinee's trait level is estimated during test administration, and items appropriate to the examinee are selected from an item bank. Items are selected to match the examinee's estimated ability according to the item response theory model that is assumed to describe an examinee's response behaviour (p.18).

In terms of a computerized adaptive test procedure, Wise (1997, p.2) mentioned that:

> It is basically a two-step process. At step one, an item is chosen whose difficulty is matched to examinee's current or initial proficiency estimate. At the next step, the examinee's response to the administered item is scored and the examinee's proficiency estimate is updated. These two steps are then repeated until some stopping criterion is met, which is usually a predetermined number of items or a desired level of measurement precision. By this process, the Computerized Adaptive Test algorithm converges on a final proficiency estimate for the examinee (p.2).

Previous research has shown that there are many benefits of computerized adaptive testing, such as enhanced measurement precision, testing on demand, reduced time for testing, reduced test length and increased testing efficiency (Wainer, 1993; Wise, 1997) which are, in part, achieved through the application of a maximum-information item-selection algorithm. Such an algorithm is an effective means for tailoring the difficulty of adaptive tests to examinee ability. The precision of

measurement of a computerized adaptive test is based on Item Response Theory (IRT) methodologies (Birnbaum, 1968; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Lord, 1970, 1971a, 1977, 1980; Lord & Novick, 1968; Rasch, 1980/1960). Item Response Theory postulates that examinees differ in their abilities on a unidimensional continuum ranging from low to high abilities. For each examinee, the probability of answering each item correctly is dependent on the current ability estimate of the examinee. It is a function of the information that is provided by the individual items in the overall item pool. When the pools are limited in their composition, the results are less ideal.

Using an item bank is one alternative method to solving assessment problems. Because a lot of good items are collected in item banks, "item banks are files of various suitable test items that are coded by subject area, instructional level, instructional objective measured, and various pertinent item characteristics (result as item difficulty and discriminating power)"(Gronlund, 1998, p.130). According to Rudner (1998a; 1998b), there are three advantages of an item bank over normal tests. The first is the development of high quality test items. The second is the test developer can "deposit" or "withdraw" items as needed. The third is it helps establish a language for discussing curriculum goals and objectives.

Instructors can use a computer to assist them in creating a computerized item bank. There are other benefits of a computerized item bank. They are; improved test quality, test security, easy to develop items to be standardized, quick and ready to construct, and accurately targeted at examinees as needed, easy to make content and statistically parallel test forms, saving in testing time and a safe area to collect the test. Moreover, a computerized item bank can encourage instructors to construct and develop test items to be standardized for use in schools.

Mathematics courses for schooling in Thailand are very important for all educational levels. All students should include mathematics in their studies during the basic education year (Ministry of Education, 2001). In higher education levels, mathematics is one of the core courses and so it is important for all students in Thai schooling. Thai educators realize the importance of mathematics, for example, in terms of daily living, in civic life, and in working. Moreover, mathematics is a necessary basic skill for studying many other subjects at school and in life.

Many students have problems that cause low mathematics achievement. Problems arise from the curriculum, teaching, learning, and assessing (Meejang & Poonpun, 1999). Assessing in mathematics is often a problem in Thai schools, for example, most tests are only used once, test items are often of poor quality, the items are not matched to the abilities of the students because the items are analysed by using Classical Test Theory (True Score Theory). One significant problem that causes low mathematics achievement is students often have negative attitudes to Mathematics. Most researchers, teachers and educators try to develop the new innovations to solve the problems (Meejang & Poonpun, 1999). The researcher is interested in using information technology in mathematics testing, by creating computer programs for Mathematics Adaptive Testing, and an attitude questionnaire to the Mathematics Computerized Adaptive Testing for primary schools in Thailand.

The Purpose of the Study

The aims of this study were to:

1. Construct multiple choice test items for a Mathematics course on the topic of Equations for Year 6 (Prathom Suksa 6) students in Thailand;
2. Calibrate the test items using a Rasch Measurement Model, investigate the validity and reliability of the test data, and set up the items in an item bank of Computerized Adaptive Testing;
3. Create a computer program for Computerized Adaptive Testing;
4. Test the program and modify the program as appropriate ;
5. Construct and develop an attitude questionnaire to the Mathematics Computerized Adaptive Testing;
6. Investigate the Mathematics abilities and attitudes to the Mathematics Computerized Adaptive Testing of Thailand Year 6 students;
7. Compare the test length, testing time, and mathematics competency for different stopping criteria;
8. Compare the test length and testing time among differences in mathematics competency of the examinees; and

Research Questions

1. Can the difficulties of the items in the 'bank' be modeled and aligned on a scale of Mathematics achievement from easy to hard using a Rasch measurement model?
2. Can the Mathematics Computerized Adaptive Testing software be used to examine differences in mathematics competency of Year 6 students in Thailand?
3. Are changes in test length and testing times related to different stopping criteria in Computerized Adaptive Testing?
4. Are changes in test length and testing times related to differences in mathematics competency of the examinees?
5. Can the attitude to the Mathematics Computerized Adaptive Testing of the Year 6 students in Thailand be measured using a Rasch measurement model and aligned from low to high on the same scale?
6. What are Thailand Year 6 student abilities in mathematics and attitudes to the Mathematics Computerized Adaptive Testing?
7. Are there changes in measured mathematics ability using Computerized Adaptive Testing when different stopping criteria are applied?

The Significance of the Study

In this study, the researcher needs to solve some mathematics assessment problems by constructing and calibrating the mathematics test items using the then latest computer program RUMM (Rasch Unidimensional Measurement Model)(Andrich, Sheridan, & Luo, 2003) that has not been used in Thailand for this before. The research set up the items in a computerized item bank for Computerized Adaptive Testing, created a new computer program using Computerized Adaptive Testing suitable for

primary schools in Thailand, created an attitude questionnaire for Mathematics Computerized Adaptive Testing, and calibrated the attitude items by using the computer program-RUMM. This study will help the teachers and students in primary school in Thailand to:

1. Have a new computerized item bank for the Computerized Adaptive testing consisting of multiple choice test items for mathematics which are calibrated by using the RUMM computer program (Andrich et al., 2003). The program is considered the most recent and probably the best computer software for analysing data with a 1-parameter Rasch measurement model.

2. Have a new computer program using an interesting method, Computerized Adaptive Testing. This method is adaptable to different student abilities, and hopefully, will help the students perform the test with care, interest and enthusiasm.

3. Have tests which are tailored to the abilities of examinees and which can reduce testing time and test length.

4. Design and implement a new attitude questionnaire to the Mathematics Computerized Adaptive Testing suit for primary schools in Thailand.

Definition of Terms

Computerized Adaptive Testing (CAT) refers to testing with items which are appropriate for the examinees' abilities, administered by students using a computer. The selection of each item depends on the result of the answer of a previous item.

Computerized Item bank (CIB) refers to a group of efficient items which are constructed according to the proper principles of testing construction, and the constructed items are to be systematically kept and administered by using a computer. In this thesis, the items in the bank were Year 6 mathematics items, each item was calibrated by using the RUMM program (Andrich et al., 2003) and all items fitted the measurement model.

Item selection refers to the method of selection of each item from the item pool by matching the item difficulty and mathematics ability of the examinee which is estimated by the CAT program..

Stopping criteria (SC) refers to the values specified for stopping the testing of each examinee. The standard error of estimation (SEE) of an examinee's ability will be used. In the present thesis, four types of stopping criteria, $\text{SEE} \leq 0.20$, $\text{SEE} \leq 0.30$, $\text{SEE} \leq 0.40$, and $SEE_m - SEE_{m-1} \leq .005$, were used.

Ability estimation method refers to the method used to estimate an examinee ability using the Computerized Adaptive Testing process. In the present research, an Updating Bayesian method was used to calculate because an examinee ability estimated using this method is more stable, least-biased, and more accurate than the others, when there are less than 500 examinees in testing (Skaggs & Stevenson, 1989; Weiss & McBride, 1984).

Attitude towards Mathematics Computerized Adaptive Testing refers to opinions of student to The Mathematics Computerized Adaptive Testing procedure.

Structure of the Thesis

This thesis is reported in nine chapters.

Chapter two describes a review of the relevant literature, including Item Banks, Computerized Adaptive Testing and Attitude towards Computerized Adaptive Testing.

Chapter three presents the theoretical framework of the study. It explains the development of items for a small 'bank' and Computerized Adaptive Testing (CAT). The theory of Computerized Adaptive Testing involving terminal criteria, examinee ability, and how the CAT works in theory are explained.

Chapter four describes the measurement of the study. The chapter starts with a description of problems with current classical measurement approaches, before the Rasch measurement model is introduced. The Rasch measurement model solves these problems and is used to analyse a new attitude and behaviour measurement questionnaire and mathematics test that is used in the present study. Requirements and equations for the

Simple Logistic Model of Rasch and the Partial Credit Model of Rasch are then provided, followed by some important outputs of the RUMM computer program.

Chapter five presents the research methodology of the study. This chapter explains the ethics and administrative procedures used. The details of the item bank construction are then explained, as well as Computerized Adaptive Testing (CAT), and attitude measures. Administrative approaches and ethics details and problems are also explained. Item bank construction concerns the mathematics items, piloting testing, data collection, student samples and data analysis. Computerized Adaptive Testing involves student samples, piloting testing, data collection, and data analysis.

Chapter six describes the process of data analysis for the mathematics item bank, using the Rasch Unidimensional Measurement Model (RUMM) computer program. The item locations, residual, chi-square, and probability of the items fitting a model are presented. A discussion of the validity and reliability of the test are also provided in this chapter.

Chapter seven contains a description of the results for the computerized adaptive testing, using a computer program designed by the author. The SPSS computer program was used to analyse data. The frequencies, percentages, and one way ANOVA were used to examine the results.

Chapter eight reports the data analysis of attitude questionnaire to the mathematics computerized adaptive testing of students using RUMM (2010).

Chapter nine answers the research questions as well as implications for relevant persons and implication for future research.

The next chapter is the literature review.

CHAPTER 2

LITERATURE REVIEW

This chapter discusses the concept of an item bank for school subjects, Computerized Adaptive Testing for school subjects and attitudes towards Computerized Adaptive Testing. The rationale, practices and the problems in their development and management are presented within the context of what other researchers and educators have published on these issues.

Item Banks

Introduction to the Need for Item Banks

Item banks are potentially very helpful for teachers and test developers. The idea of item banking is associated with the need for making test construction easier, faster and more efficient. In the United States, for example, the concept of item banking has been associated with the movements to both individualized instruction and behavioural objectives in the 1960s (Hambleton, 1986; Umar, 1999). Van der Linden (1986 cited in Umar, 1999) viewed item banking as a new practice in test development, as a product of the introduction of Item Response Theory (IRT), and the extensive use of computers in modem society. Therefore, when a large collection of good items is available to either teachers or test developers, much of the burden of test construction can be removed. The quality of tests used in the schools, for example, could be expected to be better than it could be without an item bank. When a calibrated item bank is developed under IRT, testing programs can be made more flexible and appropriate, because different groups of students can take different tests which are suitable to each of them and the results can still be compared on the same scale.

Traditional assessment (as in True Score Theory) and its tools cause many problems in education, such as, a circular dependency: (a) the person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item

difficulty and item discrimination) are (examinee) sample dependent (Fan, 1998, pp. 357-381), the items were not conceptualised in order from easy to hard, the theoretical ordering of item difficulties is not tested with the 'real' data to create a linear scale and the item difficulties (from easy to hard) and the person measures (from low to high) are not calibrated on the same interval-level scale (see also Chapter four for the problems).

Rasch measurement coupled with item banking has the potential to overcome some of these problems. It is possible to produce high quality items that not only ensure more accuracy in evaluating learning achievement but also provide an alternative way to enhance the educational system as a whole. Item banking coupled with Rasch measurement could result in improvements in school learning and in school reporting of achievement (Umar, 1999).

> A large collection of good items will help teachers to concentrate more on their teaching without having to spend much time on item construction. It could also ensure that only high quality items are used. When such a collection (popularly referred to as an "item bank") consists of items measuring the same thing and calibrated onto a common scale, it could help test developers in solving many of the practical testing problems. Use of a calibrated item bank could thus affect policies in educational testing and assessment (Umar, 1999, p.207).

What is an Item Bank?

Generally, the words item banks and item pools are used interchangeably in the research literature. Scholars generally identify the term, Item Bank, as a large collection of good test items for which their quality is analysed and known, and which are systematically stored in a 'bank' and accessible to students for measuring their achievement or ability (Choppin, 1981; Department of Academics, 1991; Millman & Arter, 1984, pp.315-316; Paeratkool, 1975). The items can be stored and retrieved by different aspects, such as subject area, instructional objective measurement, measurement traits, and significant item statistics such as item difficulty and discriminating power. The item bank is intended to ease the search and application of various testing procedures and to serve the users' needs (Department of Academics, 1991, p.4; Gronlund, 1998, p.130).

Some scholars state that item collection is not only a 'warehouse' or 'storage house' of items but, in a proper item bank, the items are systematically organized

through the processes from the start. In a proper item bank, each of the items is codified and classified by subject matter assessed, objectives, and the psychometric traits of the items. The well-selected items are normally stored in the memory unit of the computer so that they can be later easily used when needed (Ebel & Frisbie, 1986, p.927). Ideally, the advancement of item banking could be achieved in that the statistical processes will be applied to differentiate and aggregate the items with the same difficulty level. This contributes to the possibility of the assessment comparison, although the results are gained from different test items (Shoemaker, 1976 cited in Lila, 1996, p.36; Wright & Bell, 1984, p.331).

The concept of item banking can be divided into two categories: conventional and 'temporary' (van der Linden,1994 cited in Srisamran, 1997, p.7). In a conventional item bank, there is standardization of the items, their construction and their storage. An emphasis is placed on experimental control consisting of four components. One, a test blueprint table of specifications (or a two dimensional table) is constructed to indicate the relationship between the subject matter being tested and the behavioural objectives needing to be measured. It indicates the test's content validity. Two, test items are created in accordance with the table of specifications. Three, then the following procedures are performed: (1) measurement of each item's quality in regard to accuracy, objectivity, index of item of content and objective congruence by experts; (2) The item and its overall test are then analysed based upon model of Classical Test Theory (True Score Theory) in order to seek its item difficulty, its discriminating power, and the reliability of the test (Lord, 1980, p.8). Four, the investigation of norms are performed in order to compare and interpret the scores obtained with the common standardized scores.

In 'temporary' item bank, a new paradigm of test construction has been derived and test item banking has been developed with the application of statistics. Each test item is statistically calibrated to be on the same scale on the basis of Item Response Theory and Rasch measurement (see also Chapters Three and Four). This can be easily processed with a specially developed computer program which in turn produces each item of the test that fits the measurement model (see also Chapter Four). The test is therefore made more flexible and appropriate by the new concept and its implementation. This has been explained by van der Linden (1986 cited in Umar, 1999, p.209) who viewed item banking as a new practice in test development, as a product of

the introduction of Item Response Theory, Rasch measurement and the extensive use of computers in modern society In item banking, the items which cover every aspect of the domains are categorised and stored into the same domain of knowledge or ability. They are also located on a common scale. In the selection of the items for testing, such as Computerized Adaptive Testing, a certain statistical value namely difficulty is considered to be appropriate for the ability or competence level of the student. The result of the test even though different items are used can be compared since each of the test items is on a common, calculated linear scale. Hence, item banking of a calibrated item bank can not be separated from Item Response Theory itself. An item bank at this level could be considered as a model of a 'measurement system'. In this system, any new items intended for measuring the same attribute could be validated and calibrated onto the existing scale of the bank. Since the items are calibrated, it is possible to compare results from tests consisting of different subsets, of items from the bank (Hambleton, Sawaminathan, & Rogers, 1991). As such, a calibrated item bank when developed under Item Response Theory makes the testing programs more flexible and appropriate, because different groups of students can take different items which are suitable to each of them and the results can still be compared on the same scale. Together with sophisticated computer software, application of Computerized Adaptive Testing could be made possible at the school or district level (Hambleton et al., 1991).

Potential Benefits of Item Banking

With regard to the benefits of item banking, it is believed that item banking can potentially bring several advantages to educational assessment. The students could directly benefit from such an evaluation tool since the well-developed test items can potentially accurately predict their true competence or achievement level. There are ten potential benefits of item banking gleaned from the literature.

(1) The teachers can select good test items which meet the measurement objectives and the content from the item bank to suit their students' abilities in each of the area of testing.

(2) The item banking can reduce time spent on the construction of the test items by teachers. This could result in teachers having more time available for the students and their teaching tasks (Umar, 1990).

(3) The items analyzed using Rasch measurement will help create a test which contains items located on a common, linear scale and based on a variety of options or objectives (Rudner, 1998a) which in turn contribute to the comparison of the test results of the students who take the different test items since the Rasch model used, will assure items from multiple tests can be placed on a common scale and indicate the relative difficulty of the items (Rudner, 1998a).

(4) The item bank will enable teachers to build a test which contains items located on a common, linear scale and based on a variety of options or objectives by using the Rasch measurement model which is highly effective in item analysis and unidimensionality assessment (Njiru & Romanoski, 2007, pp.3-4; Rudner, 1998a, 1998b).

(5) Item Banking displays the advancement and standards in a school's measurements of student achievement.

(6) Teachers and measurement experts will be able to easily improve the item bank either by increasing or improving the test items to make them updated and relevant to the changing curriculum, as is required by State Systems, schools and the public at school and national levels (Njiru & Romanoski, 2007, pp.3-4).

(7) A well-developed item banking enhances effective measurements because the test items can be improved in both validity and reliability to meet educational higher standards (Umar, 1990). This consequently assures the accuracy and reliability of the measurement.

(8) Security is guaranteed because there are a lot of items in the bank. It is unlikely that the students who take the test can remember all of the items from one or several testings. Item banks can therefore protect item leakage, at least to a large extent (Choppin,1981 cited in Millman & Arter, 1984; Umar, 1999, p.210).

(9) Item banking is a product of a new innovation in measurement, namely Rasch measurement coupled with improvements in computing power (Computerized Adaptive

Testing), and is easily applied to school state and national educational assessment; each student can complete different test items but the results from the testing can be compared (Umar, 1999).

(10) Item banking potentially allows for the creation of a test which is adaptive to any group of students who have different learning abilities and for students with disabilities (Umar, 1990).

Limitations of Item Banks

Although these two types of item banks are an improvement on existing assessment methods, they do have some limitations and restrictions. For example, the test constructed is fixed both in terms of content and items. Additionally, when the curriculum and the content are developed or changed, it consequently influences the validity of the test, if used again. The flexibility of the test is also problematic, since it cannot be again used with the same group of the test takers. Also, in the case where the competence of the students varies greatly, the measures gained from the test can vary greatly from the likely true scores (Lord, 1980, p.8; Lord & Novick, 1968).

Item banking involves equating various tests and items. It is entirely possible, mathematically, to equate tests which cover entirely different subject matter. At the practical level, this means that it is also possible to equate items which assess subtly, but significantly different skills. In order to avoid this undesirable situation, the item review process must also include a careful evaluation of the skills assessed by each item and tests must be carefully formulated (Lawrence, 1998; Njiru & Romanoski, 2007).

While it is possible for a school or state to implement very successful item banks and Rasch-calibrated testing programs without knowing anything about IRT, good practice calls for a staff that is comfortable with, and knowledgeable of, what they are doing. A school or state that decides to undertake an item banking project should have full understanding of the practical as well as the mathematical/theoretical aspects of item banking (Lawrence, 1998; Njiru & Romanoski, 2007).

An item bank really consists of multiple collections of items with fairly unidimensional content area, such as mathematic computations or vocabulary. In order to develop the bank, many tests must be calibrated, linked (or equated), and organized. This requires a great deal of work in terms of preparation and planning and in terms of computer time and expertise. Once the item bank is established, however, test development time, effort, and cost are reduced (Lawrence, 1998; Njiru & Romanoski, 2007).

Conclusively, it can be seen that most of the problems on Item Banking are technical-practical problems (Njiru & Romanoski, 2007). Hiscox (1983) pointed out that it is not all that easy to implement several aspects of a successful item bank, such as securing or developing a sound and useful collection of items, having knowledgeable people to maintain the item bank, publicizing the item bank, and using the items appropriately and effectively. Some of these concerns, however, apply to tests constructed by traditional means as well (Njiru & Romanoski, 2007).

Item Banking in Thailand

In the case of Thailand, the concept of item banking apparently emerged in 1957 and was widely known in 1982-1984 when Thailand was assigned by her neighbouring Asean countries to initiate a testing program for the entire Asean education, but its use in any Asian country is very limited, probably because of the large cost involved in development (Boonprasert, 1988). Throughout the 1982-1984 project, there were several training seminars and further educational seminars, including the proceedings for the meetings. Since then the Thai Ministry of Education has been slowly developing item banking with a view to eventually expanding it to the regional and local levels (Department of Academics, 1991, p.5). At the Provincial level, for example, the Item Banking and Examination Online System Chiang Mai Examination Center was established in Chiang Mai Province in 2007 (Sangphueng & Chooprateep, 2007), and the Project of Item Banking Development of Nong Khai Superintendents was established in 1997 (Srisamran, 1997), but these have not been developed to the stage where they can be used by teachers and students in schools on a continual basis. They are still in the developing and trailing stage.

On Thai university campuses, there has been some limited research of item banking such as the Online Test Bank at Sura Nari University of Technology (Chansilp, 2006). The test items in this university were standardized on the basis of Traditional Measurement Theory which can only produce non-linear scores and so it is difficult to see how this item bank project can be useful and it would have been better if the researchers had used Item Response Measurement Theory to create linear measures. Other item bank projects in Thai universities have used Item Response Measurement, but they have used the now discredited so-called 2-parameter model (actually involving three parameters, item difficulty, item discrimination and one parameter of person ability) or the so-called 3-parameter model (actually involving four parameters, item difficulty, item discrimination, a guessing parameter and one parameter of person ability) (see Wright, 1999b for a disscussion and discrediting of these models). The best Rasch model to use is the so-called 1-parameter model (actually one parameter of item difficulty and one parameter of person ability) (see Andrich, 1988a, 1988b; Wright, 1999b). In Thailand, the 2-parameter and 3-parameter models were used by instructers and research students to develop trials of item banks for Mathematics (Maneelek, 1997; Songsang, 2004; Supeesut, 1998; Tuntavanitch, 2006), English (Phungkham, 1988), and Chemistry (Suwannoi, 1989).

Item Banking in Other Countries

Some studies on item banking using Rasch measurement models in different subject areas have been evident in some countries over two decades. For instance, Gerhon (1990) in vocabulary, Westers & Kekderman (1990) in mathematics, Nakamura (2001) in languge, and Njiru & Romanoski (2007) in Physics. Njiru and Romanoski (2007) developed and calibrated Physics items from the Tertiary Entrance Examination (TEE) in Western Australia. They employed the Rasch measurement model in the calibration of the 1997-2006 Physics items, using the computer program Rasch Unidimensional Measurement Models (RUMM 2020), created by Andrich, Sheridan and Luo (2005). Through the process they used 174 items that fitted the model to install them in the item bank. Based on their findings, they suggested to teachers that the item bank can be utilized in a variety of multi-purposes. Teachers, for example, might use the items in the bank to design a class assessment, diagnose students' needs, or determine achievement levels.

Nakamura (2001) created a multiple choice language test and employed the one-parameter Rasch model in analyzing the test items. He found that Item Response Theory

and the Rasch measurement model introduced a new approach to language test development that allowed examiners to adjust a constructed test by adding or removing some items from the bank, without reducing the accuracy of the measurement.

Computerized Adaptive Testing

The Meaning of Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) uses a computer to select items in a test in which items are initially selected from a bank of items. The test items are constructed and calibrated, and items of more appropriate difficulty to the ability level of the individual test-takers are selected after each choice (Beevers, McGuire, Stirling, & Wild, 1995; Lord, 1971a, 1980; Nering, 1996; Shermis et al., 1996; Stocking & Swanson, 1998, p.271; Wainer, 1990; Weiss & Kingsbury, 1984). CAT consists of an optimally informative set of items given a particular person (Embreston & Reise, 2000; Weiss, 2004). Examinees do not have to answer exactly the same test items as any other examinees and the number of test items to be answered by different examinees are not equal, they depend on the result of the test items that an examinee chooses to answer (Karnjanawasri, 2002; Lord, 1980; Weiss & Kingsbury, 1984).

Advantages of Computerized Adaptive Testing

There has been some limited research in the area of Computerized Adaptive Testing conducted over the last 20 years. The researchers stress that Computerized Adaptive Testing is more efficient than conventional paper-and-pencil tests, because the questions in a computerized adaptive test are tailored to an individual examinee's ability level. Computerized Adaptive Testing also offers advantages to tests developers in regards to improved test reliability, improved test security and data collection, better opportunity to control cheating, and cost saving with regard to printing and shipping. Convenience and flexibility of scheduling an appointment to test, anytime testing, immediacy in test scoring and reporting, faster score reporting service, potentially shorter tests, reduced scheduling and supervision, fewer test items to arrive at a more accurate estimate of test-taker proficiency levels, and reduction of teacher time on marking are also the advantages of CAT (Green, 1984; Karnjanawasri, 2002; Leung Chi Keung, 2001; Meijer & Nering, 1999; Owen, 1975; Patsula & Steffen, 1997; Wainer, 1990; Weiss, 1982; Wright & Masters, 1982). Different tests can also be equated and combined for use in an item bank (Sadeghi & Tognolini, 2006).

Some scholars such as Meijer and Nering (1999), credit CAT with various benefits which it has over traditional testing or paper and pencil tests. Enhanced measurement precision, and testing on demand also make CAT very useful for attractive and less length. In terms of reduction of test length, Shermis, Stemmer, and Webb (1996) conducted the pilot study of CAT in the Michigan Educational Assessment Program (MEAP). Over 500 volunteer students in grade 9-12 answered 97 items by using five paper and pencil forms of mathematical content. Each form consisted of 23-25 items. All forms consisted of six similar items. The computerized adaptive version (HYPERCAT) was used by 122 volunteers in a different group. The data from paper and pencil forms were calibrated and vertically equated by using RASCAL. They found that CAT could reduce test length by 25%. They also found that the CAT version assessed student achievement better than the paper and pencil form. Moreover, examinations based on CAT can achieve at least as good precision as a paper-and-pencil test, using only half of the number of items (Embreston & Reise, 2000; Weiss, 2004). However, the initial costs of implementing and launching CAT are high. Considerable financial and human resources are needed to staff and organized a CAT program. In many cases, complicated technical, economic, and political changes are also needed (Sands, Waters, & McBride, 1997). For example, although test security initially seemed to be one of the greatest advantages of CAT, it became one of its major problems. Item banks needed to be continually updated to ensure item and test security. This greatly increased the cost of implementing an operational CAT. Although CAT applications do have certain problems, their advantages outweigh their disadvantages (Meijer & Nering, 1999).

In addition, CAT also offes a mathematical programming approach that creates a model that take can care of many questions concerning the test, such as feasibility, accuracy and time of testing, as well as item pool security (Cordova & Mario, 1998). CAT could be used to obtain the most information about a single test taker compared to paper and pencil tests including methods for estimating an examinee's ability, based on the (dichotomous) responses to the items in the test.

Psychologically, CAT helps lessen the stress of the test-taker since those with lower ability do not have to do tests that are too difficult for them or too long. This makes CAT goes hand in hand with the fact that each student is challenged at his or her own level because items that are too difficult or too easy for a given student need not be

administered (Eggen & Verschoor, 2006; Karnjanawasri, 2002; van der Linden & Pashley, 2000).

In terms of the reliability, validity, fairness and feasibility, adaptive testing takes advantage of technology and modern measurement theory to deliver tests that are more reliable. Since only items of appropriate difficulty are administered to test takers, lower measurement error and higher reliability can be achieved using fewer items. When items are targeted to the ability level of the examinee, the standard error of measure (SEM) is minimized and test length can be minimized without loss of precision. Thus CAT can substantially reduce test length compared to paper and pencil tests (Gershon, 2005; J B Olsen, Maynes, Slawson, & Ho, 1986; Weiss, 1983; Weiss & Kingsbury, 1984).

CAT helps ensure that: (1) the test measures what it purports to measure; (2) the inferences made from the test scores are meaningful and useful, and; (3) the content of the test reflects critical aspects of the crucial skills or knowledge. Shorter tests with acceptable precision, possible with CAT, can enhance validity when examinee fatigue or test anxiety may introduce construct irrelevant variance (Gershon, 2005, p.112; Gershon & Bergstrom, 1995; Huff & Sireci, 2001).

Computer adaptive tests also have characteristics that enhance fairness. Since tests are administered via the computer from a large bank of items, there is no human intervention on the selection of test forms. Given the existence of a well constructed item bank, each test taker has the same opportunity to demonstrate ability or achievement as any other test taker. Recent improvements in electronic test publishing ensure that banks can be swapped easily in and out allowing compromised items to be removed from circulation in real time (Gershon, 2005, p.113).

From a cost perspective, adaptive tests are feasible for many organizations. The cost for administering adaptive tests is spread out over several areas that roughly conform to the test development and administration cost structure of any exam at a comparable level of security: test content development, test administration, scoring and reporting. Test content development for CAT differs in terms of the number of items required to create an item bank large enough to cover the range of abilities, and also large enough to insure overall bank security. For criterion referenced mastery tests, the

test may only need to have a large number of items near a pass point, but for a norm referenced test, a large number of items may be required across the ability or trait continuum. For high stakes tests, administered to thousands of examinees, it may be necessary to have a large number of items to merely insure test security. At the other extreme are low stakes and/or self-assessment tests where a very small bank of less than 100 items may be sufficient (Gershon, 2005).

The cost consideration for item development is primarily of concern for high stakes norm-referenced testing programs. Once items have been written, the next cost relates to calibrating the item response theory parameters for every item. In the case of an established testing program using previously administered items, the calculation of bank parameters may simply require re-analysing old data sets. At the other extreme, all newly written items may have to be piloted on hundreds of examinees. While it is clear that many organizations will experience increased up-front costs to create their CAT program, they may similarly encounter decreased costs in the future, as the necessity to write completely new tests each year is replaced by lesser bank maintenance tasks such as insuring the currency of existing items (getting rid of items that are now outdated), and writing a greatly reduced number of new items each year to insure content coverage and to further increase security by keeping the bank fresh.

The cost of test administration is also related to the security level of the test. High stakes tests must be administered in proctored settings. Third- party test delivery vendors, with test administration centers located in thousands of cities throughout the United States and around the world, act as sub-contractors to provide a secure high stakes test environment. Alternatively, a test administration organization can set up its own private centers on a full-time or part time basis. Lower stakes CAT exams can now be administered over the Internet. While the testing time for a CAT is typically shorter than its fixed length test equivalent, test administration time at a testing vendor is often paid for based upon the maximum time allotted for testing. The cost of scoring a CAT is basically nonexistent, since the scoring burden is born in the test administration process itself. There are no bubble sheets to collect and scan, and indeed, for many organizations, the final score report is produced on screen or on paper at the time of testing; removing the cost of generating reports altogether (Gershon, 2005, p.113).

History of Computerized Adaptive Testing

The history of Computerized Adaptive Testing (CAT) can be traced back to 1960s, when there had been the development of the Rasch model and Item Response Theory (IRT) (Lord, 1952; Rasch, 1960; Wright & Stone, 1979). The two notions have provided a theoretical structure for building large scale calibrated item banks (Choppin, 1985). One of the first adaptive tests to be developed was the ASVAB (Armed Services Vocational Aptitude Battery). The stimulus for producing a CAT test for personal selection and classification in the Armed Services was to increase the accuracy of test scores, reduce test compromise and reduce testing time. The first conference of CAT researchers for the ASVAB, held in 1975, was followed by several years of research. In addition to designing the test, the Navy Personnel Research and Development Center (NPRDC) researchers designed a complete delivery system (Gershon, 2005).

In 1979, computer technology was simply not ready to address CAT- ASVAB requirements. Much of the early effort by NPRDC and Service researchers served as a learning experience, while they waited for computer hardware to catch up with the functional requirements of the CAT-ASVAB (Gershon, 2005, p.27).

From 1979 to 1992, the NPRDC researched, developed, tested and implemented several generations of the CAT-AS VAB. By the mid-1980's experimental CAT-ASVAB data from over 7,500 military recruits from all Services had been collected and analysed. The CAT-ASVAB system remained in operational use until 1996 when it was replaced by the 'next generation' system (Gershon, 2005, p.112).

This includes several other works of scholars in the late 70s to the early 90s. A meta- analysis of 20 studies published from 1977 to 1992 compared results from paper and pencil administrations to CAT administrations, and consistently found that both modes of test administration yielded similar results (Bergstrom and Lunz, 1992 cited in Gershon, 2005, p.112). English, Reckase, and Patience (1977 cited in Gershon, 2005, p.112) published a study of undergraduate students enrolled in a course entitled "Introduction to Educational Measurement and Evaluation" at the University of Missouri. Bejar and Weiss (1978 cited in Gershon, 2005, p.112) reported on achievement test results for students enrolled in a large introductory biology class at the University of Minnesota. The California Assessment Program used mathematics application items to create tests in a pencil and paper administered format, a computer

administered format, and a computer adaptive format (Olsen, Maynes, Slawson, and Ho, 1986 cited in Gershon, 2005, p.112). Comparability of CAT and pencil and paper versions of the mathematics computation section of the College Level Academic Skills Test (CLAST) at the University of Florida were reported by Legg and Buhr (1987 cited in Gershon, 2005, p.112). The results of computer administered and pencil and paper versions of the Differential Aptitude Test, a battery of eight ability tests, were reported by Henly, Klebe, McBride and Cudeck (1989). Baghi, Gabrys and Ferrara (1992 cited in Gershon, 2005, p.112) conducted research done with the Maryland Functional Testing Program, a state wide competency testing program used as a high school graduation requirement. The study compared paper-and-pencil versions and computer adaptive versions of mathematics and reading tests and illustrated the previously mentioned issues with long text reading passages. Both the American Society of Clinical Pathologists (Gershon, 2005; Lunz & Bergstrom, 1991) and the National Council State Boards of Nursing (Gershon, 2005, p.112) reported on studies that demonstrated the validity of CAT.

When looking closer, we can see that each of these studies (despite differences intent content, age of test-takers, latent trait model (Rasch or IRT) used and study design) demonstrated the comparability of measures obtained using CAT and pencil and paper test versions. Indeed, what is most remarkable in reviewing the literature comparing these two test modalities is the marked absence of any significant studies demonstrating the *inability of* CAT to capture measure originally assessed using paper tests. Even the minor decrement in performance realized with long reading passages in CAT, may in reality prove that the CAT format better captures reading comprehension. The paper format may benefit the test taker who is quick to re-scan the material, and the CAT version may benefit the examinee who is better able to commit the material to memory (Gershon, 2005, p.112).

Types of Adaptive Testing

There are two main types of adaptive testing: (1) "two-stage strategies" and (2) "multi-stage strategies", where the classifications are based on strategies in item selection (Hambleton & Swaminathan, 1985; Weiss, 1974).

Two-stage strategies involve a test that is adapted to suit the examinees' proficiency level by providing two steps of testing. The first step, or routing test, generally consists of 10 items aiming at identifying the examinees' ability. The result,

or the examinee's ability, will be used in selecting a suitable sub-test in the second step. In the second step, a main test or a measurement test that consists of many sub-tests is provided. The sub-tests range from easy to difficult. Each sub-test contains 20-30 items. An examinee who achieves a high result from the first step will take a difficult sub-test in the second step. The medium and low ability examinees will take medium and easy sub-tests, accordingly (see details in Lord, 1971c; Weiss & Betz, 1973). Many educators, including Linn, Rock and Cleary (1969) and Lord (1971a; Lord, 1980), found that using "two-stage" adaptive testing helps decrease test length without reducing accuracy of the measurement. However, the utilization of a test with a large number of examinees requires immediate scoring on the routing test and this requires a lot of scorers. If an error occurs at stage one (in the routing testing), there will be an error in a selection of the measurement test at stage two. Consequently, an opportunity to make errors in classifying the examinees' competency can be up to 20 % (Weiss & Betz, 1973).

Multi-stage strategies involve the selection of items in response to each of the previous items, in the form of a "Branching Tree" (Thissen & Mislevy, 1990, p.110). The test contains many items of different levels of difficulty. Usually, a more difficult item for the correct response will be chosen as a next item. Practically, the initial item is moderate, not too easy or too difficult for most examinees. If the response for the initial item is correct, the next item will be more difficult. In cases where the examinee gives an incorrect answer, the next item will be easier. The test ends when the examinee meets the stopping criterion. There are two types of multi-stage strategies, Fixed-branching and Variable-branching. These strategies differ in their structure, item ordering, item selection and the stopping criteria.

Generally, a Fixed-branching strategy in a test is adapted to suit each examinee and contains many stages. Each stage consists of one item or more. A certain line of response is set in advance. Many models of Fixed-branching strategy have been evident, i.e., "Constant Step Size Pyramidal Testing" (Sukamolson, 1996), "Variable Step Size Pyramidal Testing" (Lord, 1971a, p.93), Robin-Monro model (Lord, 1971a, p.95) "Truncated Pyramidal Testing" (Mussio, 1973 cited in Weiss, 1974, p.102), "Multi-item Pyramidal Testing" (Krathwohl and Huyser, 1956 Linn, 1969 cited in Weiss, 1974, p.105), and "Differential Response Option Branching Pyramidal Testing" (Bayroff and Seeley, 1968 cited in Weiss, 1974, p.109), "Flexi-level Test" (Lord, 1971b) and "Stradaptive Test" (Water and Bayroff, 1971 cited in Sukamolson, 1996, p.48).

Unlike the Fixed-branching strategy, a Variable-branching strategy does not set item and line of responses in advance. Instead of scoring the previous response, a Variable-branching strategy applies examinee's competence estimation after each response. Then, the next item is chosen in response to the previous item. According to the estimation for each response, it is inconvenient to use paper-and-pencil adaptive tests. Therefore, the computer is introduced to operate the test, provide estimations and record results. The test begins when an examinee sits in front of a computer screen, processes personal information required in the test, and reads the test instructions. An initial item will be presented on the screen. After each response the next item will be popped up, one at a time, in according to the item selection criteria. This will continue until the test achieves the stopping criteria. Finally, the test result will be presented on the screen (Songsaeng, 2004).

The present study applies Multi-stage strategies focusing on the Variable-branching model for its convenience in selecting the items, scoring, estimating the examinees' competence, and for conducting the test and providing accuracy in measurement. Also, it does not require a pre-selection of the number of items and lines of responses. In addition, the researcher chose standard error of estimation as a stopping criterion.

Attitudes Towards Computerized Adaptive Testing

Positive attitudes towards CAT have been evident in Thailand and other countries. The evidence is divided into sections of positive attitudes across different disciplines, positive attitudes towards different types of CAT, preference for CAT over traditional tests and some negative attitudes towards CAT.

Positive Attitudes Across Disciplines

There has been evidence showing positive attitudes towards CAT across disciplines and at all levels. Baghi, Gabrys, and Ferrara (1991) conducted a five year research study, from academic year 1985 to 1990, on the applications of computer-adaptive testing with mathematics and reading in Maryland, USA. The subjects were the eighth and ninth graders in 24 school districts. The students' attitudes towards the CAT were very favorable and positive on the CAT-Math test and the CAT-Reading test respectively. The students were positive towards the clarity of the test directions, sample

items at the beginning of the test-taking procedures and the clarity of the item graphics. Kenyon and Malabonga (2001) examined attitudinal reactions to taking different formats of oral proficiency assessments across three languages: Spanish, Arabic and Chinese. Participants were graduate and undergraduate students taking language courses at their universities. It was found that the Computerized Oral Proficiency Instrument (COPI) allowed the difficulty level of the assessment to mathematics to be more appropriate to the proficiency level of the examinee. These examinees reported that the COPI helped lessen the test difficulty.

Similar findings were found in Thailand. Songsang (2004) found that the 135 sixth graders had very satisfactory attitudes towards CAT. Supeesut (1998) found that Thai seventh graders were pleased with the clarity of the items, symbols and the test itself. They preferred the CAT to its interesting, immediate feedback, easy to give and change answers, and its free-from-worry to do the test. They could finish the test shortly. They could do the items that suit their proficiency level and were willing to keep trying on the difficult ones. They reported relatively little worries in doing the CAT.

Moreover, Suwannoi (1989) reported that the eleventh graders who participated in CAT in Chemistry at the Demonstration School, Modindaeng, Khon Kaen university, Thailand had positive attitudes towards the CAT. The students paid attention and were interesting in, and willing to do, the test. They reported that the CAT encouraged their perseverance and lessened their anxiety. Sukamolson (1996) found that the first year undergraduates at Chulalongkorn university, Thailand, were satisfied with the CAT-English test and were motivated to complete the test.

Positive Attitudes Towards All Types of CAT

Regarding types of the CAT were employed and reported very satisfactory. For instance, the Pyramidal Testing was employed by Suwannoi (1989), and Sukamolasan (1996) and got high satisfaction from the examinees for its conveniences.

The Bayesian strategy was administered in different assessments (for example, Pomsit, 2001; Songsaeng, 2004; Supeesut, 1998) and found that it was at high satisfactory. Pomsit (2001) revealed that the Bayesian strategy on web page satisfied the examinees for its immediate feedback, easy to proceed data and to give answers. These encouraged the examinees' interest and lessened their anxiety. They were positive for clarity of the test directions and procedures.

The Computerized Two Stage Test was conducted by La-ongkaew (1995) with the fifth graders at the Demonstration School, Modindaeng, Khon Kaen university, Thailand and found that the students had positive opinions to the test. They were motivated to do the test and the test helped lessen their frustration.

Preference of CAT to Traditional Tests

Compared with traditional tests it is found that CAT is more preferable (see for example, Baghi et al., 1991; Kenyon & Malabonga, 2001; Pomsit, 2001; Sukamolsan, 1996; Suwannoi, 1989; Vicino & Moreno, 1997). Vicinio and Moreno (1997) were in line with Baghi and others (1991) in that the test-takers preferred the computerized test over paper and pencil test. They mentioned less or no longer anxious with the CAT than with the paper and pencil test. Also they no more faced a difficulty in reading on the screen compared to the test booklet.

Similarly, Suwannoi (1989), Sukamolason (1996) and Pomsit (2001) found that the Thai examinees had more interest and motivation in using CAT than doing the booklet test. In addition, Pomsit (2001) found that, if there were choices, students preferred taking a CAT to traditional tests.

Some Negative Attitudes Towards CAT

Some negative attitudes towards CAT were also reported. For instance, Baghi, Ferrara and Gabrys (1992) found that the ninth graders were bothered by the inability to change their answers after pressing the enter key. Those who failed the reading test in their study indicated greater problems in scrolling through the reading paragraphs and thought that reading a paragraph on the screen was more difficult than reading from a booklet. Students who had never used the computer reported problems in using the space bar and scrolling through the reading paragraphs. Sukamolson (1996) also found that the CAT was complicated and costly to create and conduct. Meanwhile, Vicina and Moreno (1997) found a problem that many test-takers faced in their study of CAT use, that is, not being able to review and modify answers to previous questions.

Based on the literature review in this chapter, a theoretical framework involved around the use and development of CAT was set up. It is presented in the next chapter, Chapter Three.

CHAPTER 3

THEORETICAL FRAMEWORK

This chapter explains the development of items for a small 'bank' and Computerized Adaptive Testing (CAT). The theory of Computerized Adaptive Testing involving terminal criteria, examinee ability, how the CAT works in theory are explained.

Reasons for Developing Items for a Small Item Bank

Educators are now encountering an apparent gap existing between the modern and traditional methods of measurements. Many important tests have been constructed or revised by measurement principles that differ qualitatively from classical measurement concepts. True-score theory, for example, uses test items that are developed with all items having approximately the same difficulty (Waugh & Chapman, 2005). Item analysis is described by the characteristics of inter-item correlations and item discrimination. The items, as a result, are not conceptualised in order from easy to hard and, therefore, are not appropriate for use in an item bank.

In a contrast, new measurement principles like those involved with Rasch measurement provide the examinees with items that are particularly informative about their abilities levels because the items are ordered by difficulty level. Different examinees, as a consequence, can take different ability tests which are more suited to their abilities. The new tests can be now computerized and administered in an adaptive form as part of an Item Response Theory. Item Response Theory has many practical advantages for test development. Unlike classical test theory, in Item Response Theory item parameters are not biased by the population ability distribution whereas, in classical test theory, the indices for item difficulty and discrimination are directly influenced by ability distributions. Furthermore, greater flexibility in test calibration, using item subsets with varying groups, is possible because Item Response Theory readily handles missing data problems (Embretson & Hershberger, 1999, p.vii).

In all phases of test development, Item Response Theory plays an important role because it is the method that makes adaptive testing practically feasible. In cognitive ability testing, computerization is the most salient change in the new generation of tests. Computerized presentation of items, immediate scoring, and report generation are attractive features of many revised tests. Computerized testing also has made adaptive testing feasible. In adaptive testing, tests no longer have fixed-item content. Items are selected online for an examinee, depending on their responses to proceeding items. Thus, examinees no longer are exposed to items that are far above or below their performance level. Test forms are optimally selected for each person from the test item bank. Another salient change in cognitive ability testing is increased flexibility for administering and interpreting individualized tests, such as the Differential Ability Scales (Elliot, 1990; Woodcock & Johnson, 1977), and several others. Special procedures for missing data in testing (such as, persons measured out of level or omitted items) are available so that ability may be estimated without bias. Furthermore, some individual cognitive tests also provide ability estimates that do not depend on a norm-referenced standard for meaning. The ability estimates have optimal scale properties that permit comparisons directly to abilities obtained earlier or to abilities at another developmental level. The abilities may be used to measure developmental change or distance from some developmental standard.

Furthermore, Item Response Theory has important applications in calibrating items and measuring individual abilities. Explicating the nature of the latent constructs underlying performance, establishing the applicability of the constructs to varying groups of people (such as racial-ethnic groups, gender groups, clinical populations, non-native speakers), and establishing scalability are important issues in construct development (Embretson & Hershberger, 1999, p.viii). Item Response Theory is increasingly employed in construct development due to its many advantages over classical test theory approaches.

In all, Item Response Theory is very useful in the construct development phase of testing. It now includes a vast array of models that postulate qualitatively different types of underlying constructs. Comparative fit indices for different Item Response Theory models can provide interpretations about the constructs that are measured. For example, inconsistent findings about the number and nature of constructs involved in specific tests result, in part, from applying methods that are inappropriate for item-level data. Applying multidimensional Item Response Theory models to item level data

results in more valid findings. Furthermore, it is often suspected that some test items are population-specific; that is, performance may differ qualitatively over different groups of persons. Sometimes the populations are intrinsic to the measure, such as employing different strategies to solve the items. Other times the populations differ in background, such as defined by gender, racial-ethnic background, native language, or clinical status (handicaps, disabilities.). Item Response Theory models are available not only to assess these differences, but their application can provide solutions.

In Item Response Theory models, the probability of a response to a test item is the result of an interaction between the properties of the item and the trait level (or ability) of an examinee. This interaction is typically mapped on the parameters of the examinee and the item. One of the main advantages of separate parameterizations is that it is possible to select items to match the trait level or ability of the examinees. A standard approach to assembling a conventional linear test is to select a combination of items from an item bank with optimal values for their information functions over the interval of the scale in which the examinees are expected to be (Birnbaum, 1968). A more powerful application of the principle is found in computerized adaptive testing, in which each item in the test is selected to match the current estimate of examinee ability (trait) (Wainer, 1990).

The measurement theory used for the present study is an attempt to apply Item Response Theory using a Rasch measurement model. Item Response Theory is based on the notion of a relationship between the observable responses to test items and the unobservable traits assumed to underlie responses to items on a test. A mathematical formula is used to describe this relationship (Hambleton & Swaminathan, 1985; Rasch, 1980/1960). Item Response Theory is a family of mathematical models that describe how people interact with *test* items (Andrich, 1988b; Embreston & Reise, 2000). These models were originally developed for *test* items that are scored dichotomously (correct or incorrect), but the concepts and methods of Item Response Theory extend to a wide variety of polytomous models for all types of psychological variables that are measured by rating scales of various kinds (van der Linden & Hambleton, 1997). These ideas are applied to the construction of an item bank for mathematical equations accessed using a new computerized adaptive testing program by the researcher.

In the construction of test items on mathematical equations for Prathom Suksa 6 students in Thailand, general learning achievement principles focusing on constructing the items in accordance with learning objectives and the coverage of content taught

were taken into consideration. This includes six behavioural objectives. They are: (1) Given several symbolic sentences, students can identify the equation; (2) Given several equations, students can identify the true equation; (3) Given several equations, students can identify the equation with an unknown identity; (4) Given an equation with unknown identity, students can choose the number and substitute the unknown identity; (5) Given the equation with the unknown identity on addition, subtraction, multiplication and division, students can tell how to find the solution, and solve the equation correctly; and (6) Given the problems relating to daily life which require addition, subtraction, multiplication and division, students can convert the problems into the equation, and solve it to get the answer.

There were nine aspects relating to the equations. They are: (1) identification of equations from given choices; (2) identification of the true equation; (3) identification of an equation with an unknown; (4) finding the true equation in different circumstances; (5) finding the method to solve the equations; (6) finding the solution of an equation; (7) finding the solution or equation which related to the given conditions, (8) selecting an equation which is converted from a verbal problem or a problem which is converted from an equation; and (9) problem solving.

Model of the Structure of Mathematics Items

A model of the structure of mathematics achievement on equations was conceptualised nine the nine main aspects of achievement (mentioned above). The test items are created in an ordered pattern by difficulty within each aspect. The structure of achievement was then based on sub-sets of test items in patterns of ordered difficulty, each aligned from easy to hard (see Chapter Six). This involved calibrating all the difficulties of the items (from easy to hard) onto the same scale as the measure of mathematics achievement (from low to high), using a Rasch Measurement Model. The following material provides an example of the conceptual thinking involved with the construction of one aspect, selecting an equation converted from a daily problem, or a problem converted from an equation.

Example

Expected ordering by difficulty pattern for selecting an equation which is converted from a verbal problem or a verbal problem which is converted from an equation

It was expected that most students would find it very easy to select an equation which is converted from a verbal problem " Y" students in a classroom were divided into 8 equal groups with 5 students each" (item 1). It was expected that most students would find it harder (but still easy) to select an equation to find out the value of X from a problem "Dang had X Baht and had 10 Baht more from selling eggs. The total sum of his money was 30 Baht was ???" (item 2). It was expected that students would find it harder again (but still easy) to choose an equation which shows how many pieces of paper Pooh collected from a problem "Pooh had 3 pieces of paper and she collected Z pieces more. The total pieces were 20" (item 3). The equation in Item 3 requires students to figure out the difference between the two pieces of information and therefore, is more challenging than item 2, which is only a simple addition equation. Item 1 is easiest because it only requires students to create an equation by converting the given problem directly.

It was expected that they would find it moderately hard to select an equation which shows the total sum of John from a problem "John had the sum Y Baht. He bought a flashlight for 120 Baht and two bags for 70 Baht, and 55 Baht remains" (item 4). It was expected that students would find it harder still to create an equation which shows how many pieces Adam bought from a problem "Adam bought Z pieces of pork, costing 3 Baht per each. The sum used was 54 Baht." (item 5). In creating an equation for Item 4, students need to write up an equation for finding the difference between three things. The more things there are, the higher-level thinking is required. An equation in response to Item 5 involves multiplication, which is more challenging than subtraction and addition equations.

It was expected that most students would find it hard to select a verbal problem, relating the equation: $X \div 5 = 7$ (item 6). Rather than providing a verbal problem for converting to an equation as in items 1-5, this item requires students to find out the verbal problem representing the given equation. This requires students to think thoroughly and critically, and they need to try five different equations to find out the

correct equation. In addition, a division equation requires higher-level thinking than does the equation involving multiplication, subtraction and addition.

The vertical ordering of test items by difficulty is set out in Table 3.1.

Expected Ordering by Difficulty for the Other Aspects

The test-items for the other aspects were designed to be ordered vertically from easy to hard. The actual reasoning is not reported here to avoid repetition, but it can easily be worked out from Tables 6.2-6.10 in Chapter Six.

Table 3.1
Conceptual order of difficulty of equations involving conversion from a verbal problem or a verbal problem which is converted from an equation

Item Number	Item content	Difficulty
1	Select an equation of the statement " Y students in a classroom was divided in to 8 equal groups with 5 students each.	very easy
2	Select an equation in finding out the value of X from a problem "Dang had X Baht and had 10 Baht more from selling eggs. The total sum of his money was 30 Baht."	easy
3	Select an equation, which shows how many pieces of paper did Pooh collect from a problem "Pooh had 3 pieces of paper. She collected Z pieces more. The total pieces were 20".	easy still but harder
4	Select an equation, which shows the total sum of John from a problem "John had the sum Y Baht. He bought a flashlight for 120 Baht and two bags for 70 Baht. 55 Baht remains."	moderately hard
5	Select an equation, which shows how many pieces, did Adam buy from a problem "Adam bought Z pieces of pork, costing 3 Baht per each. The sum used was 54 Baht."	still more moderately hard
6	Select a verbal problem which is related the equation $X \div 5 = 7$	hard

Notes on Table 3.1

1. Items are designed to be ordered by perspective from easy to hard (vertical ordering).
2. Source: part of the test designed by the researcher for this study.

The Theory of Computerized Adaptive Testing

Computerized Adaptive Testing is derived from the notion of computer-based testing. Computer-based testing is a form of assessment which is applicable for both high stakes tests such as certification or licensure examinations, as well as health-related quality of life surveys. Computer based testing was initially implemented on mainframe systems dating back to the early sixties. But its use was primarily limited to the military and some large corporate and private training companies who could afford to purchase their own hardware (Gershon & Bergstrom, 1995 cited in Gershon, 2005, p.111).

Computers are now readily available in practically every setting, high-speed access is easily attainable and the Internet has become a basic component of many facets of daily living. Today, computers are very common, while computer-based testing is a special accomplishment, that may become universal in the future. Computerized Adaptive Testing refers to a form of computer-based test administration in which each test-taker takes a 'customized' or 'tailored' test. Test taker competence is assessed after each item is administered and the next item is targeted to the current estimate of ability (Gershon & Bergstrom,1995 cited in Gershon, 2005, p.111).

The advantages of Computerized Adaptive Testing, besides its general accessibility and basic metrics for measurement, include reliability, validity, fairness and feasibility. Adaptive testing is more reliable than the usual paper and pencil tests, because, besides taking advantage of technology and modern measurement theory to deliver tests, only items of appropriate difficulty are administered to the test-taker, there is lower measurement error, and higher reliability can be achieved using fewer items. Furthermore, Computerized Adaptive Testing can also reduce test length compared to paper and pencil tests because items are targeted to the ability level of the examinee, the standard error of measure is minimized and the test length can also be minimized without loss of precision (Olsen, Maynes, Slawson, & Ho, 1986; Weiss, 1983; Weiss & Kingsbury, 1984). Shorter tests with acceptable precision of Computerized Adaptive Testing can enhance validity (in comparison to traditional tests) because examinee exhaustion and test anxiety which may introduce construct irrelevant variance can be reduced. Complicated item selection algorithms built into Computerized Adaptive Testing can ensure that content is balanced for each test-taker (Gershon & Bergstrom, 1991 cited in Gershon, 2005, p.112; Huff & Sireci, 2001). Administration of the test

through the computer from a large bank of items and no human intervention on the selection of test forms promotes fairness in Computerized Adaptive Testing. Through a well-constructed item bank, each examinee has the same opportunity to display ability or achievement as any other test-taker. Some advantages of Computerized Adaptive Testing involve inexpensive test administration, better test content development, easier scoring and reporting compared to the other conventional test types (Gershon, 2005).

The development of the Rasch model in the 1960s (Rasch, 1960; Wright & Stone, 1979) together with Item Response Theory models (Lord, 1952) provided a theoretical structure for building large scale calibrated item banks (Choppin, 1985). The ability to order all of the items on the same scale, which is important, has been attempted in the present study. That is, all items are calibrated on the same scale in an item bank, and the particular items constructed on the same content that are administered to a given test-taker become a matter of indifference. Each individualized adaptive test created from the calibrated bank is automatically equated to every other test that has been or might be drawn from the bank (Master & Evans, 1986; Wright & Bell, 1984). When items are calibrated on the same scale a pass/fail point (criterion-referenced standard) can be established for the entire item bank and thus test-takers are measured against the same criterion-referenced standard regardless of the group of test-takers with whom they are examined, the particular set of items which are administered or when they take the test (Gershon & Bergstrom, 1995 cited in Gershon, 2005, p.111).

Numerous measurement models can be used for adaptive testing, including the Rasch dichotomous model (Rasch, 1960), the 1, 2 and 3-parameter models (Hambleton et al., 1991) and the rating scale and partial credit models (Andrich, 1978; Bock, 1972; Wright & Masters, 1982) have been used for adaptive testing. For this present study the Rasch 1-parameter model has been used because it guarantees a unidimensional, linear scale, if the data fit the model, and it has been shown to be more valid and reliable than the other parameter Rasch models (Wright, 1999b).

The Rasch model is expected to work well for adaptive tests and it has been used extensively in constructing linear, unidimensional measures. In Rasch measurement models, the underlying construct, or latent trait, described by the items on a test is a continuous variable extending to negative and positive infinity on an abstract continuum. All possible test item difficulties and all possible test-taker ability levels lie on this continuum. The measure estimated for a test taker on a set of items is the result

of the interaction between the ability of the test taker and the difficulty of the items administered (Gershon, 2005, p.114).

There are two key elements at the heart of a well-developed Computerized Adaptive Testing system. The first is a large bank of accurately calibrated items that cover a wide range of difficulties. The second is a test item presentation algorithm that determines the next item to be presented on the computer screen for the current test taker. Both of these elements derive considerable benefits from the application of Rasch measurement. The construction of calibrated item banks that provide for the presentation of almost unlimited versions of person-specific tests is a consequence unique to Rasch measured Computerized Adaptive Testing. Moreover, when the intent is to operationalize another benefit of Rasch measurement, the item-selection algorithm is constrained to present items at the 50% probability of success for the current test taker, based on the success or failure on the current item. The presentations might follow, say, a 0.2 logits increase in difficulty with a successful response or a similar decrease in difficulty following an incorrect response to keep the future items well targeted for the respondent (Bond & Fox, 2001). Logits are the commonly used Rasch measurement units defined as the log odds of successfully answering an item.

In practice, the measurement of the construct is bounded by the range of measures obtainable, given the range of calibrated items administered on the test. The idea of a single continuous scale for test takers and items implies that there is a point where the ability of the test taker equals the difficulty of the item, a point where the difference between the estimate of ability and item difficulty is as close to zero as possible. Although this point can only be approximated in practice, the idea is crucial to Computerized Adaptive Testing (Gershon, 2005, p.114).

The Algorithm and Its Main Features of Computerized Adaptive Testing

A computer algorithm employed in the Computerized Adaptive Testing helps the test-taker choose items administered to each examinee. Although there have been several Computerized Adaptive Testing item selection algorithms proposed, each of these methods essentially matches item difficulty to the proficiency level of the examinee. The computer algorithm's selection of an item for administration is based on the examinee's responses to previously administered items during the testing session. This results in an efficient test administration system in which examinees are

administered different sets of items, which yield maximum information about each individual examinee. Increased testing efficiency is a primary advantage of Computerized Adaptive Testing (Wainer, 1993).

Computerized Adaptive Testing algorithms 'target' the difficulty of the test to the current ability estimate of the test taker by attempting to present an item at the point where the difference between test taker ability and item difficulty is zero. By targeting the difficulty of the items to the ability of the test taker, Computerized Adaptive Testing maximizes the information from each item so that an item that is too easy or too hard is not administered. When test information is maximized, the standard error of measure is minimized. Thus, administering items adaptively can reduce test length and improve measurement precision. If the item is more difficult than the test-taker's ability, the Rasch model predicts that the test-taker will have less than a 50% probability of correctly answering the item; if the test-taker's proficiency exceeds the difficulty of the item, the Rasch model predicts that the test-taker will have a greater than 50% probability of correctly answering the item (Bergstrom & Lunz, 1999; Wright & Stone, 1979 cited in Gershon, 2005, p.p.114-115).

According to Gershon (2005, p.114), the basic process of administering a computer adaptive test is very similar to that of conducting a simple binary search. For example, if we are asked to think of a number between one and one-hundred, a typical linear testing process would have to ask up to 100 questions to determine the correct answer. "Is the correct answer 1?" "Is it 2?," "Is it 3?" etc. Using a binary search, the same result can be located in only seven questions. If the correct answer is 74, the questioning would go something like this: "Is the number greater than 50" ... Yes. "Is it greater than 75". No. "Is it greater than 67?" .Yes. "Is it greater than 71? Yes. "Is it greater than 73?" Yes. "The answer is 74." By using a binary search, we never needed to ask about each of the first 50 numbers, because we immediately knew that the unknown value was greater than 50. Similarly, we didn't need to ask about each of the numbers above 75 after the second question.

In the Computerized Adaptive Testing process, each time an examinee responds to a question we are also able to converge on an estimate of a person's measure (zeroing in on their ability level). On a pass-fail test, we would typically administer the first item at the pass point. If that item is answered correctly, a harder item is administered. If answered incorrectly, an easier item is given. This process is iterated until specific

stopping conditions are met (such as testing until a specific level of measurement precision is obtained). Many testing options such as various stopping conditions will be elaborated on later in this manuscript.

For the algorithm of Computerized Adaptive Testing, Bunderson et.al. (1988, p.57) suggest four major steps. One, a preliminary estimate of ability is made for the examinee. Two, a test icon is selected and administered that will provide maximum information at the estimated ability level. The information value of the item can be calculated on-line or stored in a pre-computed information matrix. Generally, if the examinee answers an item correctly, a more difficult item is presented; if the examinee misses the item, an easier item is administered. Of all the items available, the one selected is calculated to maximize new information about that examinee, subject to constraints due to content balance and placed to control excessive exposure of certain items.

Three, the ability estimate is updated, or revised, after each item. A variety of methods have been proposed for ability estimate updating. The methods proposed include Bayesian Sequential Ability Estimation (Owen, 1969, 1975), Maximum Likelihood Ability Estimation (Birnbaum, 1968; Lord, 1977, 1980; Samejima, 1977), Expected A Posteriori Algorithm (Bock & Aitkin, 1981; Bock & Mislevy, 1982a) and Bi-Weighted Bayes estimates. The Bi-Weitghted Bayes is a robustified ability estimator (Bock & Mislevy, 1982b; Jones, 1982; Wainer & Thissen, 1987; Wainer & Wright, 1980).

Four, the testing process continues until a designated test termination criterion has been met. Typical termination criteria include a fixed number of test items, when the standard error reaches, or is less than, a specified value and when the test information function reaches or exceeds a specified value.

More recent advice of rules for Computerized Adaptive Testing procedure given by van der Linden (1999, p.142) suggest that an obvious way to select items in a Computerized Adaptive Testing procedure is to base the selection of subsequent items on the information functions of the items in the pool. Each next item could then be selected such that it has maximum information at the ability value where the examinee is estimated to be. In fact, this *maximum information* rule is the most popular item assignment rule in Computerized Adaptive Testing. It is not the only rule; an alternative is a Bayesian rule in which the next item is selected such that the expected variance of

the posterior distribution for the examinee's ability value is minimal. The two item assignment rules can only be operational if an estimate of the examinee's ability from his or her responses to previous items is obtained in realistic time.

A natural partner of the maximum information rule is *maximum-likelihood estimation* of ability. In this estimation method, the ability value that maximizes the likelihood of the response pattern obtained from the examinee is defined to be his or her ability estimate. The use of modern hardware and software gives quick estimates for the Item Response Theory models currently in use. For Owen's rule (1975), Bayesian ability estimation is the appropriate choice. The choice of the first item in a Computerized Adaptive Testing procedure is important, because considerable gain of information, and hence reduction of test length, can be obtained if the first item is not too far off target. If no prior information about the ability of the examinee is present, the best choice is to start with an item that is optimal at a (subjective) estimate of the location of the ability distribution of the population of examinees for which the test has been designed. If prior information is available, for example, in the form of information on background variables with a known regression on the ability variable, better choices are possible (van der Linden, 1999).

In Karnjanawasri's perspective (Karnjanawasri, 2002, p.177) of Computerized Adaptive Testing, the principles of adaptive testing selection for an individual test-taker depend upon his/her previous item answered. After the test-taker finishes answering the first item, estimates of the test taker's ability will immediately be done in order to select the following item for which the difficulty level suits the ability level of the examinee. If the answer is right, the next item is harder; but if the answer is wrong, the next will be easier. This process is repeatedly done until the level of the examinee's ability is reliably gained (that is low error); the test is then terminated. The procedure is shown in the Figure 3.1.

For this present study, the procedure of Computerized Adaptive Testing was divided into two parts: registration and implementation. The former part, as shown in Figure 3.2 includes functions such as the program registration, program provider's data, examinee's data and test instruction. The program provider's data displayed name of the programme, in this case, The Chaow Computerized Adaptive Test Computer Program as well as name of the researcher and the university. Chaow is an abbreviation of the

researcher's name. The examinee's data required a student's information, i.e., student code, name, grade, school and password (security system).

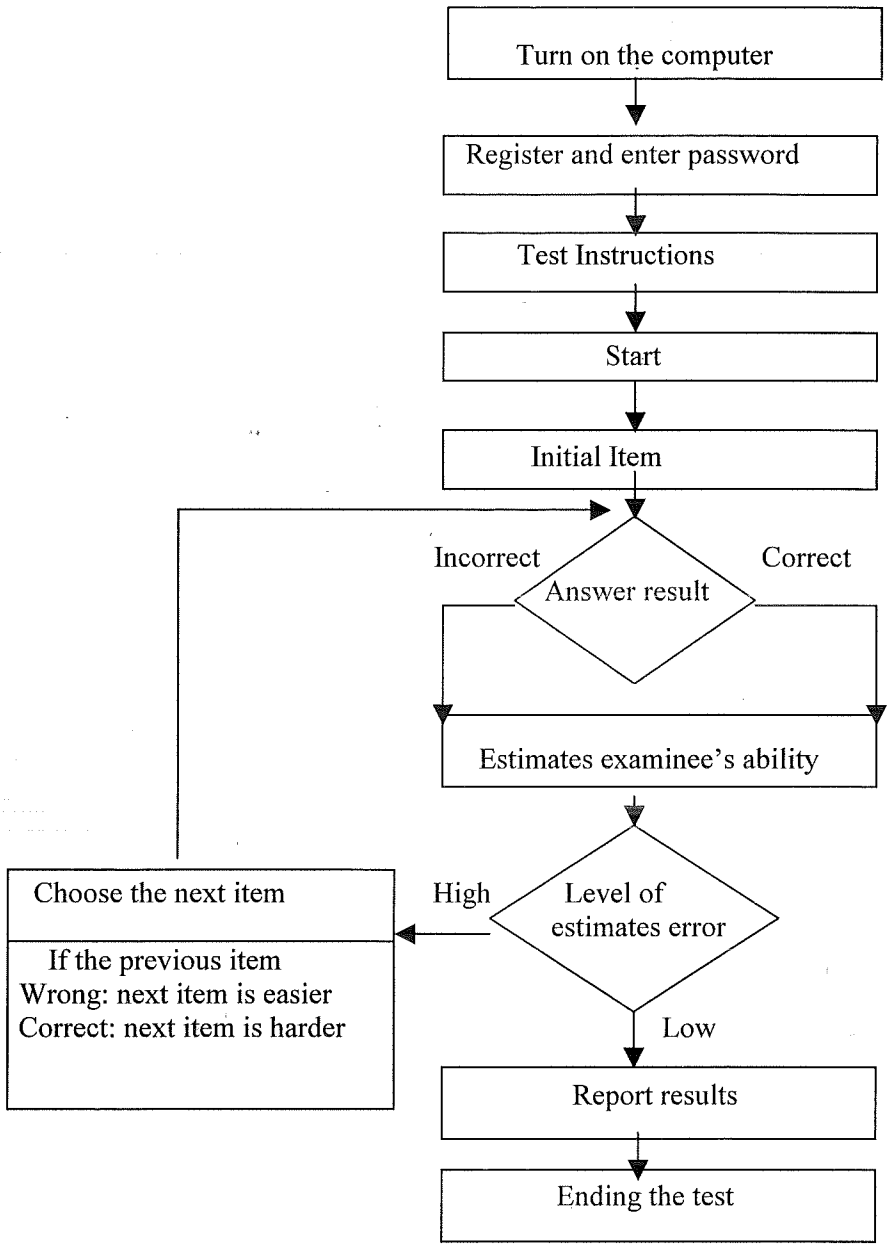


Figure 3.1 **Procedure of CAT**

Source: Karnjanawasri, 2002, p. 182 and used by the researcher for the present study.

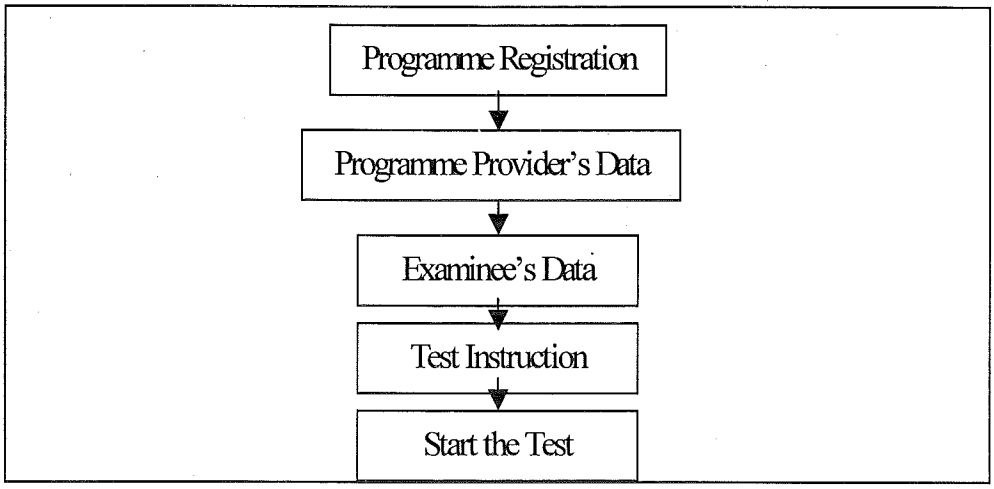


Figure 3.2 **Part I of the CAT program registration**

Source: Created by the researcher for this study.

In the latter, the implementation part starts with the initial item, which is randomised as a medium difficulty item and displayed. The choices are provided at the bottom. The examinee's response would be monitored. The correct or incorrect response would be a source for the estimation of the examinee's ability by using Bayesian Updating (Owen, 1969, 1975). According to the ability estimated (θ), the Computerized Adaptive Testing program would select and display an item with the difficulty closed to the examinee's estimated competency. Based on this response, the examinee's ability (θ) and the Standard Error of Estimation (SEE) would be calculated. The Standard Error of Estimation would be compared with the stopping criteria 1, 2, 3 or 4. If the score does not meet the criteria, another test item would pop up and the monitoring and the estimation functions would be repeated. Such interaction would keep going until the Standard Error of Estimation met the criteria. After the test ended, the test result and details would be automatically saved in the computer and reported on the screen.

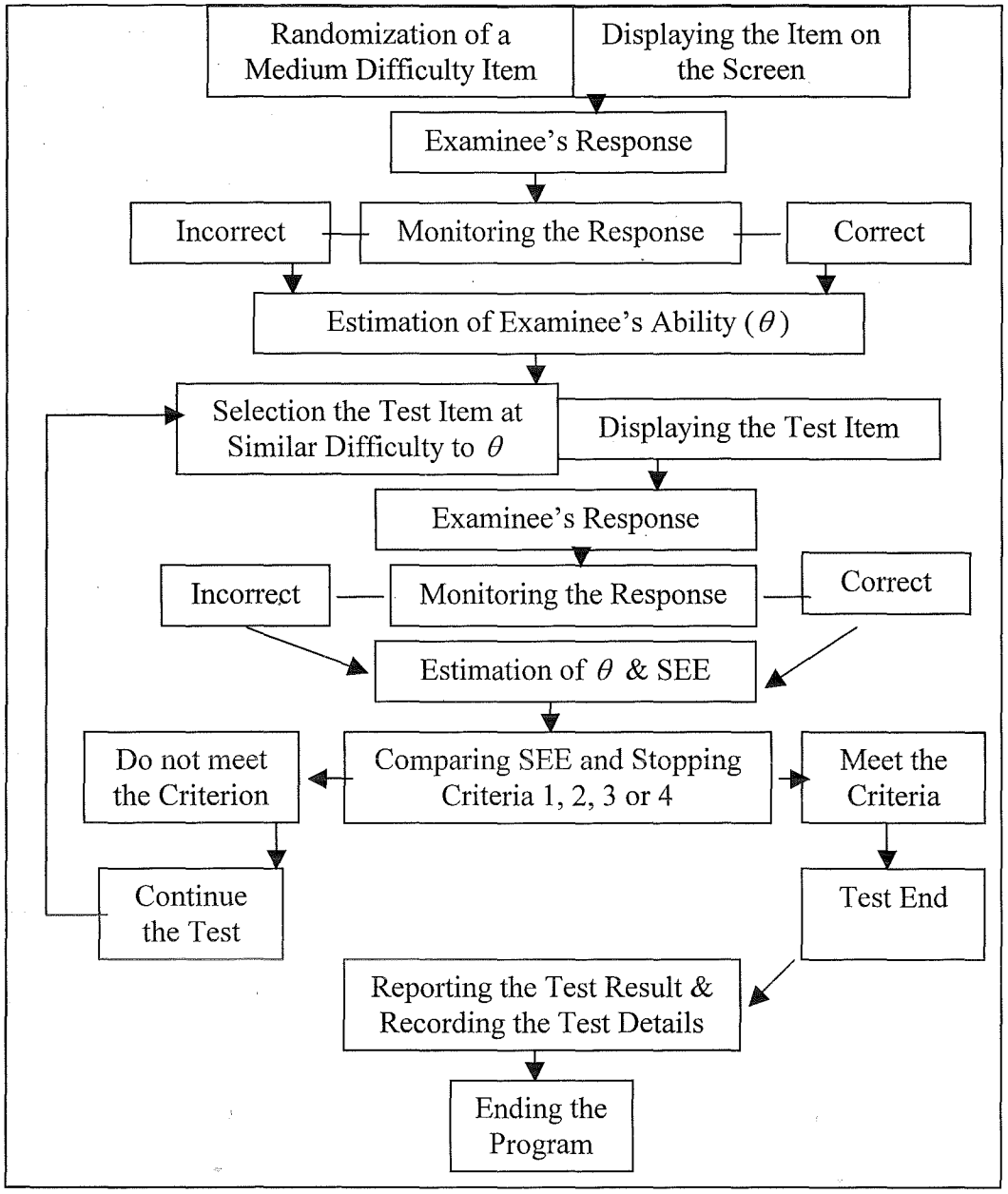


Figure 3.3 **Process of CAT**

Source: Adapted from Maneelek (1997) and Supeesut (1999).
Note: SEE = Standard Error of Estimation.

Expectation of CAT Working

It is expected that the computerized adaptive testing program created would randomly select an initial item according to the condition required by the researcher and it would also randomly select a test item from the item bank, regardless of the ability of the examinee ability. Also, the program would immediately estimate the examinee ability after the students answer each test item. The Computerized Adaptive Testing system would adaptively select an item according to the estimate of the ability of

examinee based on his or her responses to previous items. If one gets an item correct, the system would select the next item to be more difficult, and the next item would be easier if one gets a wrong answer for the one right on the screen. It was expected that Computerized Adaptive Testing would be a dynamic system that can provide tailor-made tests for individuals. This makes examinees always face items that closely match their own individually estimated ability. Furthermore, an individual test form of Computerized Adaptive Testing would be shorter as there are less inappropriate items for each individual. In terms of the termination criteria, the system would select the test to end when it reaches the criteria assigned by the researcher. The students would therefore know immediately their abilities after a stopping criteria was reached. This system is expected to be fair for individual test-takers and they would prefer this testing type to the paper and pencil test.

The Stopping Criteria for Computerized Adaptive Testing

In determining the stopping criterion for computerized adaptive testing, most researchers have used the standard error of estimation as the mean criterion. The values used by most researchers for the standard error of estimation is usually less than 0.20 (Nering, 1996), but standard errors of estimation ≤ 0.30, ≤ 0.40, and ≤ 0.50 (Maneelek, 1997) were tested in the present study. $SEE_m - SEE_{m+1} \leq 0.005$, where SEE_m is the value of the standard error of estimation of the previous item and SEE_{m+1} is the value of standard error of estimation of the current item (Supeesut, 1999).

Specifically, the current stopping criteria cover those of the following traits (Karnjanawasri, 2002, p.166). One, specify the fixed number of test items for stopping the testing of each examinee to which every individual test-taker can refer. For example, if 25 adaptive testing items are fixed, when a student finishes the 25^{th} test item, the testing is terminated. This is useful for a certain situation such as Monte Carlo Simulation because the equality of the test items promotes information functions of the test. Practically, however, this system might give a different quality in measures.

Two, specify a standard error of estimation level. In practice, CAT would be continuously taking place until the standard error of estimation of the examinee's ability (θ) reduces to meet the acceptable criteria. The testing is then finished.

According to Gershon (2005, p.119), the stopping rule for achievement tests are typically set by determining the desired level of accuracy. Testing continues until a specified standard error is reached. While this results in a variable length Computerized Adaptive Testing, test length varies less than in a stopping rule based on confidence around a cut score. A standard error type stopping rule is also used in Computerized Adaptive Testing with rating scale and partial credit models. For example, there has been an effort to research ways for physicians to assess various outcomes of medical treatments to see whether the patient progresses after numerous treatment outcome variables such *as fatigue or pain* (Gershon, 2005, p.120). In a classical test environment, gaining an accurate assessment of how a test taker is feeling might require the administration of hundreds of items across numerous dimensions. With Computerized Adaptive Testing, the same-symptom measures can be obtained in the short period of time that one sits in the waiting room prior to seeing their doctor; and the strength of a good set of rating scale or partial credit items is likely to take place in as few as 6 - 9 items, instead of the 16-75 currently used by typical paper-based surveys. Test length can be set to ensure that reliability is equal to or exceeds the reliability of the paper-and-pencil test or that confidence in pass/fail decisions is equivalent to, or exceeds, confidence obtained with paper and pencil tests.

For the present study, the standard error of estimation (SEE) of an examinee's ability is used to be the stopping criteria. There were four types of stopping criteria, $SEE_{m+1} \leq 0.20$, $SEE_{m+1} \leq 0.30$, $SEE_{m+1} \leq 0.40$, and $SEE_m - SEE_{m+1} \leq .005$. The example of stopping criteria in the program is presented in Figure 3.4.

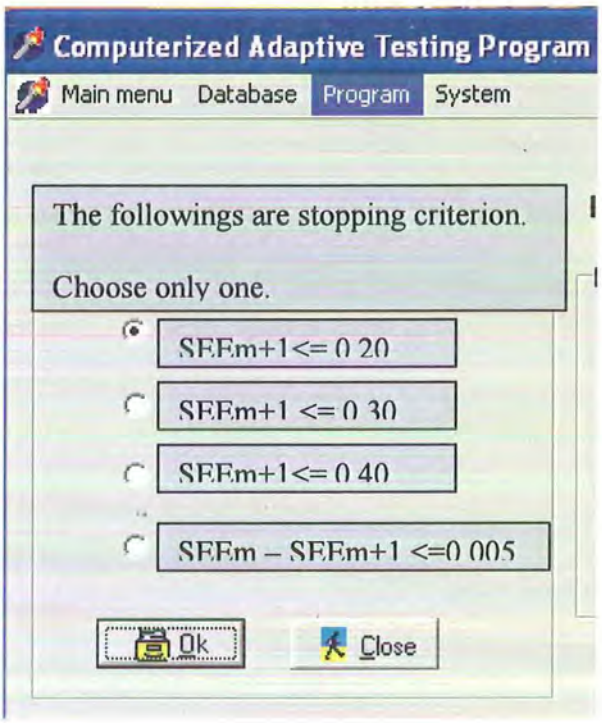


Figure 3. 4 **Four stopping criteria of the program**

Source: Created by the researcher for the present study.

Examinee's Ability

In relation to the examinee ability, there have been several methods for estimating test taker ability in Computerized Adaptive Testing. For example, in a Rasch or Item Response Theory calibrated item bank, each method examines the characteristics of the items administered, and combined with the answers given by the respondent, determines a provisional estimate of the test-takers ability. Once this estimate is obtained, the remaining items in the item pool can be evaluated in terms of the next item that will lend the greatest efficiency to the overall testing process.

Most often cited include two common methods for estimating test-taker ability, Maximum Likelihood Estimation and Bayesian estimation. In the case of Maximum Likelihood Estimation, the test-taker's ability is updated by using the difficulty of the items already administered, and the response to the most recent item. The next item selected is based on maximum information and thus the difficulty of the item selected closely matches the current estimate of test taker ability. Maximum Likelihood

Estimation with two and three-parameter models in particular is unstable for short tests and thus problematic for CAT based on these models.

The second method for estimating the ability of test takers is Bayesian Estimation. Bayesian algorithms are based on the normal distribution or on knowledge of the typical population distribution of test-takers attempting a particular examination. In the case we do not initially have any knowledge of a particular person's ability, the algorithm administers items as if the true ability estimate is close to the mean or mode of the population, updating each "prior" distribution with new information based on the test taker response (Gershon, 2005, p.115; Parshall, 2001).

In the present study, Bayesian Updating (Owen, 1969; 1975, pp.351-356) for a 1-parameter model is utilized in the estimation of the examinee's ability because an examinee ability estimated using this method is more stable, least-biased, and more accurate than the others, when there are less than 500 examinees in testing (Skaggs & Stevenson, 1989; Weiss & McBride, 1984). Its formulas for estimation are shown in the following.

In case of the examinee giving the correct answer.

$$\theta_{m+1} = \theta_m + \left\{\sigma_m^2 / \sqrt{(1+\sigma_m^2)}\right\}\{\phi(D)/\Phi(-D)\}$$

$$\sigma_{m+1}^2 = \sigma_m^2\left[1 - \left\{1/\left(1+1/\sigma_m^2\right)\right\}\{\phi(D)/\Phi(-D)\}\{\phi(D)/\Phi(\text{-D}) - \text{D}\}\right]$$

In case of the examinee giving the wrong answer.

$$\theta_{m+1} = \theta_m - \left\{\sigma_m^2 / \sqrt{(1+\sigma_m^2)}\right\}\{\phi(D)/\Phi(D)\}$$

$$\sigma_{m+1}^2 = \sigma_m^2\left[1 - \left\{\phi(D)/(1+1/\sigma_m^2)\right\}\{\phi(D)/\Phi(\text{D}) - \text{D}\}/\Phi(D)\right]$$

Where $D = (b_g - \theta_m)/\sqrt{(1+\sigma_m^2)}$

$\phi(D)$ = the ordinate value of normal curve at point D.

$\Phi(D)$ = the area under the normal curve ranging from the minimum value to point D.

θ_m = estimation of the examinee's ability before responding the test item m+1. Generally, this unknown initial estimation is set as $\theta_m = 0.00$.

σ_m^2 = the variance of the examinee's estimated ability before responding the test item m+1. Generally, this unknown initial estimation is set as $\sigma_m^2 = 1.00$.

θ_{m+1} = parameter representing the ability of the examinee estimated after the response to the test item m+1.

σ_{m+1}^2 = parameter representing the variance of the examinee's estimated ability after responding to the test item m+1.

b_g = the difficulty (location) for item m+1.

The researcher created her own computerized adaptive testing program based on the equations above, the theoretical framework explained in this chapter and the 1-parameter Rasch measurement model using the RUMM 2010 computer program developed by Andrich, Sheridan and Luo (2003). Rasch measurement procedures used in the present study are explained in the next chapter (Chapter Four) and the research method is explained in Chapter Five.

CHAPTER 4

MEASUREMENT

This chapter starts with a description of problems with current classical measurement approaches, before the Rasch measurement model is introduced. The Rasch measurement model solves these problems and is used to analyse a new attitude and behaviour measurement questionnaire and mathematics test that is used in the present study. Requirements and equations for the Simple Logistic Model of Rasch and the Partial Credit Model of Rasch are then provided, followed by some important outputs of the RUMM computer program.

Measurement

Measurement can be viewed as a process in which numbers are used to link concepts to indicators on a continuum (see Punch, 1998). Traditionally, the most common means of measuring traits and variables have been based on True Score Test theory. In True Score Test theory, item and test-analysis are described by their characteristics of inter-item correlation and item discrimination. Whereas *tests* are typically scored by counting the number of correct answers provided by an examinee, attitudes are often measured using rating scale items, by summing a set of arbitrary weights assigned to the response categories of each item (Embreston & Reise, 2000). Wright (Wright, 1999a) points out the problems with accepting True Score Theory to produce measures of educational psychology variables. Such measures are not linear and total scores on the items from such measures should not be treated as though they are linear (Wright, 1999a). Counting events does not produce equal units of measurement (Wright, 1999a), and raw, summed scores are not linear measures and shouldn't be used as though they were.

Wright (1999a) points out there are at least five problems with current classical measures, that is, with True Score Theory. One is that the items are not conceptualised in order from easy to hard. In creating a scale one must have items ordered in difficulty.

Two is that data that have only been analysed with True Score Theory cannot produce any thing better than a ranking. It certainly cannot produce a linear measure. In a linear scale, equal differences between scale numbers represent equal amounts of the variable and, students with high, medium and low measures of ability will agree that certain items are easy and that others are hard. For instance, persons with low measures are only likely to answer the easy items positively. Persons with medium level measures are likely to answer the easy and medium difficulty items, rather than the hard items, most of the time. Persons with high measures will be likely to answer all easy, medium and hard items. These characteristics are not present with True Score Theory 'measures'.

Three, in True Score Theory, item difficulties are not tested for conceptual order. That is, in True Score Theory, the theoretical ordering of item difficulties is not tested with the 'real' data to create a linear scale. The Rasch model, on the other hand, tests that item difficulties are ordered.

Four, in True Score Theory, the item difficulties (from easy to hard) and the person measures (from low to high) are not calibrated on the same interval-level scale. This is a fundamental necessity in the creation of a linear scale.

Five, in True Score Theory, the data for many measures do not show high reliability and construct validity. In the literature, there are many measures of attitude and behaviour in classrooms where reliability is 0.7 or less and where construct validity has not been adequately tested. Rasch measures, on the other hand, test for construct validity and reliability.

The measurement theory used for the present study is referred to Item Response Theory and the measurement model used is a Rasch measurement model. Item Response Theory is based on the notion of a relationship between the observable responses to test items and the unobservable traits assumed to underlie responses to items on a test. A mathematical formula is used to describe this relationship (Hambleton & Swaminathan, 1985; Rasch, 1980/1960). Item Response Theory is a family of mathematical models that describe how people interact with *test* items (Andrich, 1988a; Embreston & Reise, 2000). These models were originally developed for *test* items that are scored dichotomously (correct or incorrect), but the concepts and methods of Item Response Theory extend to a wide variety of polytomous models for all types of

psychological variables that are measured by rating scales of various kinds (van der Linden & Hambleton, 1997).

One family of measurement models based on Item Response Theory that satisfies the requirements of measurement, as suggested by Andrich (1989), is the Rasch models which have been hailed to be "simple", yet "very powerful" models of measurement (Hambleton & Swaminathan, 1985). It has also been noted that Rasch models incorporate the best elements of the Thurstone and Likert approaches (Andrich, 1982; Wright & Stone, 1979). The original Rasch model developed by Danish mathematician Georg Rasch in the 1950's, was the Simple Logistic Model (Rasch, 1980/1960), and it was used to analyse dichotomous responses. Subsequent work has extended Rasch models to incorporate polytomous responses, where three or more response categories are used to compare measures (Anderson, 1995; Andrich, 1988a, 1988b). Central to the notion of objective measurement in Rasch Models, also termed specific objectivity or sample-free measures (Andrich, 1988b; Douglas, 1982; Wright & Masters, 1982), is that both item difficulties and people measures can be calibrated on the same scale. That is, differences between pairs of person measures are scale-free and differences between pairs of item difficulties are expected to be sample-independent (Andrich, 1988b; Wright & Masters, 1982), which is a requirement of measurement.

The item difficulties used in the present study are discussed in terms of ordering from easy to hard, and calibrated on the same scale as student mathematics abilities and student attitudes, while student abilities and attitudes are ordered from low to high. Calculating item difficulties and person measures on the same scale using a Rasch measurement model will produce a linear scale. This is the reason a Rasch measurement model is used to solve measurement problems in the current study. A linear scale is better than a rank ordering and an improvement on the usual True Score Theory measures.

Rasch Measurement

A Rasch measurement model is an example of additive, conjoint, fundamental measurement. Rasch measurement models are currently the only known method by which one can create linear, objective measures applicable to the human sciences (Waugh, 2006; Wright, 1999a). Rasch measurement models show how to determine what is measurable on a linear scale, how to determine what data can be reliably used to

create a linear scale, and what data cannot be used in the creation of a linear scale. In a linear scale, equal differences between the numbers on the scale represent equal amounts of the measure. Most Rasch measures, however, do not have a true zero point, because it is difficult to know what zero achievement, attitude, personality or skill, for example, means, and so they are usually interval-level measures. That is, most Rasch scales are linear measures without a true zero point (Waugh, 2006).

Raw Scores to Linear Measures

Rasch measurement models can be used to convert many different types of raw score data to a linear scale. They can be applied to many different types of data (Waugh, 2006; Wright, 1999a). For the present study, dichotomous data from mathematics test and polytomous rating response scores from the attitude questionnaire were converted to linear scales.

'Scale-Free' Measures and 'Sample-Free' Item Difficulties

An important point to understand is that when the data fit a Rasch measurement model, the differences between the person measures and the item difficulties can be calibrated together in such a way that they are freed from the distributional properties of the incidental parameter, because of the mathematics involved in the measurement model. This means that 'scale-free' measures and 'sample-free' item difficulties can be estimated with the creation of a mathematically objective linear scale with standard units. The standard units are called logits (the log odds of successfully answering the items).

A requirement for measurement is that the units should be the same size across the range of the variable measures and this is not true with percentage scores, or summed scores from a set of achievement or attitude items, where small changes in probability of success are related to large changes in person abilities at the bottom and top of percentage scales, all of which are non-linear. By converting the probability of success to log odds and logits as the unit in Rasch measurement, the non-linear problem is greatly reduced, particularly at the top of the scale (Waugh, 2006; Wright, 1999a).

The Simple Logistic Model of Rasch

The simplest Rasch measurement model for creating a linear scale was developed by the Dane, Georg Rasch (1901-1980) and published in 1960. The Simple Logistic Model (SLM) of Rasch has two parameters: one representing a measure for each person on a variable and the other representing the difficulty for each item (it is sometimes called the one-parameter model in the literature) (Wright, 1999b).

Requirements of the SLM of Rasch

There are six requirements of the Simple Logistic Model of Rasch (Andrich, 1982; Rasch, 1980/1960; Waugh, 2006). One is that items are designed to be conceptually ordered by difficulty along an increasing continuum from easy to harder for the variable being measured.

Two is that in designing the items, one keeps in mind that person measures of the variable are conceptualised as being ordered along the continuum from low to high according to certain conditions. The conditions in this example are that persons with low measures will have a high probability of answering the easy items positively, and a low probability of answering the medium and hard items positively. Persons with medium measures will have a high probability of answering the easy and medium items positively, and a low probability of answering the hard items positively. Persons with high measures will have a high probability of answering the easy, medium and hard items positively. These conditions are tested through a Rasch analysis.

Three is that data are collected from persons on the items and scored dichotomously (0/1 or 1/2), as in, for example, but not limited to, wrong/right, no/yes, none/a lot, disagree/agree, some/often, bad/good, slow /fast.

Four is each item is represented by a number, estimated from the data that represents its difficulty (called an item parameter in the mathematical representation of the Rasch Model) that does not vary for persons with different measures of the variable. Persons with different measures responding to the items have to agree on the difficulty of the items (such as easy, medium and hard, as used in this example). If the persons do not agree on an item difficulty, then this will be indicated by a poor fit to the measurement model, and then the item may be discarded as not belonging to a measure on this continuum.

Five is that each person is represented by a number, estimated from the data that represents his or her measure of the variable (called a person parameter in the mathematical representation of the Rasch Model) that does not vary for items of different difficulty along the continuum. If different items do not produce agreement on a person measure, then this will be indicated by a poor fit to the measurement model, and then one examines the person response pattern (and the items).

The sixth is that Rasch measurement models use a probability function that allows for some variation in answering items such that, for example, a person with a high attitude measure may give a low response to an easy item, sometimes, or a person with a medium achievement measure might get a hard item right, sometimes. If the person response pattern shows too much disagreement with what is expected, then it may be that the person has not answered the items properly or consistently, and that person's results may be discarded, or the item may be too hard or too easy, requiring it to be modified. In the mathematics of the model, the probability of answering correctly is related to the difference between the person measure and the item difficulty. In situations where there is a large positive difference between the person measure and item difficulty, then there is a strong probability of a correct response and, if there is a large negative difference, then there is a strong probability of an incorrect response. If the differences are not so large, the probabilities are changed appropriately.

Equations for the Simple Logistic Model of Rasch

(Hambleton & Swaminathan, 1985; Waugh, 2006)

Probability of answering

positively (score 1) $= \dfrac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$

for person n

Probability of answering

Negatively (score 0) $= \dfrac{1}{1 + e^{(B_n - D_i)}}$

for person n

Where

e = natural logarithm base (e=2.7318)

Bn = parameter representing the measure (ability, attitude, performance) for person n

Di = parameter representing the difficulty for item i

These equations are solved from the data (entered in a text format) by taking logarithms and applying a conditional probability routine with a computer program. For the present study the Rasch Unidimensional Measurement Models (RUMM) program (Andrich et al., 2003) was used.

The Partial Credit Model of Rasch.

The Partial Credit Model (PCM) of Rasch can be thought of as an extension of the Simple Logistic Model (SLM) from two response categories to three or more response categories or outcomes. So the conditions, requirements and output of the Partial Credit Model are similar to the Simple Logistic Model, except that there are now more item parameters, more item output and the equations are more complicated. The Partial Credit Model can be applied to any set of data scored, judged or answered in three or more ordered outcome categories where the level of outcome is conceptualised on a continuum from low to high.

For the present study, the self-reported response categories are strongly disagree (score 1), disagree (score 2), neither agree nor disagree (score 3), agree (score 4) and strongly agree (score 5) on attitude towards computerized adaptive testing of Prathom Suksa 6 students. These response categories are taken to be ordered conceptually from low to high and they are scored from low to high. In this case, the Partial Credit Model gives the same result as the Rating Response Model (see Linacre, 2005; Masters, 1997).

These equations are solved from the data (entered in a text format) by taking logarithms and applying a conditional probability routine with a computer program. For the present study the Rasch Unidimensional Measurement Models (RUMM) program (Andrich et al., 2003) was used.

The RUMM computer program (Andrich et al., 2003), employing the Partial Credit Model of Rasch, has been used successfully with a number of measures in educational psychology (see for example Waugh, 2003, 2005). By successful is meant that a unidimensional measure has been obtained in which there is a good or reasonable fit to the measurement model. It has been used, amongst others, to measure attitude to mathematics (Waugh & Chapman, 2005), academic motivation (Waugh & Njiru, 2005), self-regulated learning (Njiru, 2006), university acceptance of peers with disabilities (Waugh & Biswas, 2003), teacher leadership in early childhood education (Waugh,

Boyd, & Corrie, 2003) and to attitude and behaviour to reading comprehension (Waugh, Bowering, & Torok, 2005).

Questionnaires providing data for these types of measures often use three of five response categories (such as Likert, 1932, responses). There is evidence that, conceptually and practically, the neutral category (neither agree nor disagree) is not really ordered between disagree and agree (Dubeis & Burns, 1975; Glastonbury & MacKean, 1991; Waugh, 2003; 2005). That is, for some students this neutral category may not really be neutral, but it can depend on how the students interpret it. If they interpret it as a neutral category and respond to it in that way, then the RUMM analysis could show this as four ordered thresholds for the five response categories and it will show appropriate response category curves.

Conceptually in measurement terms, neutral is not necessarily more than disagree in regard to attitude and behaviour and neutral should not necessarily be allocated a higher score than disagree. Similarly, agree is not necessarily more than neutral and so should not necessarily be allocated a higher score than neutral. Taking this line would mean that the traditional view of strongly disagree (score 1), disagree (score 2), neutral (score 3), agree (score 4), and strongly agree (score 5) is not an ordered response set (even if the numbers are ordered). The use of Likert (1932) response sets is not necessarily good measurement practice and it had to be tested in this study. The RUMM 2010 computer program provides a good test of whether the response categories (Likert or other) produce consistent and logically-used response categories in line with their conceptual construction and this was done in Chapter Eight.

Equations for the Partial Credit Model of Rasch

$$\text{Probability of person n scoring in outcome category x of item i (for x = 1,2,3,4 ...Mi)} = \frac{e^{\sum_{j=1}^{x} (Bn-\delta ij)}}{1 + \sum_{k=1}^{Mi} e^{\sum_{j=1}^{k} (Bn-\delta ij)}}$$

$$\text{Probability of person n scoring in outcome category x of item i (for x = 0)} = \frac{1}{1 + \sum_{k=1}^{Mi} e^{\sum_{j=1}^{k} (Bn-\delta ij)}}$$

Where

e = natural logarithm base (e=2.7318)

$\sum (Bn-\delta ij)$ is the sum of Bn- δij

Bn = a parameter representing the measure (ability, attitude, skill or performance) for person n

δi1, δi2, δi3,.... δiMi = are a set of parameters for item i which jointly locate the model probability curves for item i. There are Mi item parameters for an item with Mi +1 outcome categories.

Source: (Hambleton & Swaminathan, 1985; Waugh, 2006; Wright, 1999b)

The RUMM Computer Program

The RUMM computer program (Andrich et al., 2003) is currently the best of the main computer programs for Rasch measures for three reasons. One is that the RUMM program provides a comprehensive set of output data to test many aspects of both the conceptual model of the variable, the answering consistency of the response categories, both item and person fit to the measurement model, and targeting. Two is that in addition to the output data, the RUMM program produces a wonderful set of coloured, graphical maps for many aspects of the measurement. The third reason is that it has a very fast switching time from one set of output to another (Waugh, 2006).

RUMM Output

The eight data analysis tests to fit a linear scale created with the RUMM computer program output provided in the creation of a linear, uni-dimensional scale (Andrich et al., 2003; Waugh, 2006) are now given. One is testing that the response categories are answered consistently and logically. The RUMM program does this with two outputs: one, it calculates threshold values between the response categories for each item (where there are odds of 1:1 of answering in adjacent categories) and, two, it provides response category curves showing the graphical relationship between the linear measure and the probability of answering each response category.

Two is testing for dimensionality and an item-trait test-of-fit is calculated as a chi-square with a corresponding probability of fit (Andrich & van Schoubroeck, 1989). It tests the interaction between the responses to the items and the person measures along the variable and shows the collective agreement for all items across persons of different measures along the scale. If there is no significant interaction, one can infer that a single parameter can be used to describe each person's response to the different item difficulties and thus we have a uni-dimensional measure.

Three is testing for good global Item-Person Fit Statistics. The item-person test-of-fit examines the response patterns for items across persons and the person-item test-of-fit examines the response patterns for persons across items (see Styes & Andrich, 1993, p. 914 for the equations) using residuals. Residuals are the differences between the actual responses and the expected responses as estimated from the parameters of the measurement model. When these residuals are summed and standardized, they will

approximate a distribution with a mean near zero and standard deviation near one, when the data fit a Rasch measurement model.

Four is a Person Separation Index. Using the estimates of the person measures and their standard errors, the RUMM program calculates a Person Separation Index that is constructed from a ratio of the estimated true variance among person measures and the estimated observed variance among person measures. This tests whether the standard errors are much smaller than the differences between the person measures.

Five is testing for good individual item and person residuals. Residuals are the differences between the observed values and the expected values estimated from the parameters of the Rasch measurement model. It is instructive to examine these outputs as they give an indication of whether persons are answering items in a consistent way and they give an indication of individual person and individual item fit to the measurement model.

Six is Item Characteristic Curves. Item Characteristic Curves examine how well the items differentiate between persons with measures above and below the item location. It also shows a comparison between the observed and expected proportions correct for a number of class intervals of persons.

Seven is Person Measure/Item Difficulty Map. The RUMM program produces two types of person measure/item difficulty maps. These maps show how the person measures are distributed along the variable and how the item difficulties are distributed along the same variable (measured in logits). They show which items are easy, which ones are of medium difficulty and which ones are hard. They show how well the item difficulties are targeted at the person measures. That is, they show whether the items are too easy or too hard for the persons being measured and whether new items need to be added, or whether there are too many items of similar difficulty (some of which are thus not needed).

Eight is testing for construct validity. Suppose that your items are conceptually ordered by increasing difficulty (downwards) and the perspectives are ordered by increasing difficulty (to the right) and this represents the structure behind your variable. In Rasch measurement, all the item difficulties are calculated on the same linear scale and so the item difficulties can be compared with their conceptualised order. In this case, the item difficulties increase vertically downwards for each perspective by item

and they increase horizontally to the right for each item by perspective. This provides strong support for the structure of the variable as it was postulated before the data were collected and analysed.

Measuring Mathematics Achievement

The mathematics achievement test was created using nine aspects. The nine aspects are: (1) identification of an equation from given choices, (2) identification of the true equation, (3) identification of an equation with an unknown, (4) finding the true equation in different circumstances, (5) finding the method to solve the equations, (6) finding the solution of an equation, (7) finding the solution or equation which relates to the given conditions, (8) selection of an equation which is converted from a verbal problem or a problem which is converted from an equation, and (9) problem solving. The items relating to each aspect were ordered conceptually by difficulty.

They are arranged according to an order of increasing difficulty, conceptualised theoretically. This order was tested in Chapter Six. The test items were grouped under their aspect headings, so that it would be clear to the students what was being asked of them. Thus all items were written in a positive sense, with an ordered response format, from easy to hard. The tests were revised also after discussions with the researcher's supervisor. Ordering is essential for establishing a unidimensional scale as described by Waugh (2002, pp. 67-68).

All items were written in Thai. An English translation was used for discussions. The multiple choices items have four alternatives: a, b, c, and d. There is only one correct answer. The correct answer gets one point, and the wrong one gets zero. The whole mathematics items are given in Appendix K and some items are given in Table 4.1.

Table 4.1
Part of mathematics achievement test

Item Number	Test item
	Aspect: Identification of an equation from given choices
1	Which choice is the equation? a. $4X + 5 \neq 9$ b. $3B - 7 = 10$ c. $6M + 15 > 20$ d. $P - 2 \leq 4$
2	In which choice are both statements equations? a. $72 \neq 10 + 20$, $50 + 70 = 100 + 20$ b. $120-30 < 100-5$, $110+10 \neq 220 \div 2$ c. $360=36 \times 10$, $79 + 35 = 35 + 79$ d. $125 \div 5 > 23$, $1000 \div 4 \neq 200$
	Aspect: Identification of the true equation
3	Which choice is true? a. $750 \div 6 = 6 \div 750$ b. $(60 + 12) - 10 = 72 - 10$ c. $8 \times (22 \times 23) = (8 \times 22) + (8 \times 23)$ d. $3,960 = 3,000 + 90 + 600$
	Aspect: Identification of an equation with an unknown
4	Which choice has an unknown? a. $120X + 5 = 245$ b. $24 + 10 = 10 + 10 + 14$ c. $47 - 16 = 30 - 0$ d. $58 \times 5 = 5 \times 58$
5	Which choice has an unknown in both equations? a. $30 \times 3 = 90$, $X + 10 = 30$ b. $60P - 20 = 40$, $39Q \div 13 = 21$ c. $79 \times 35 = 35 \times 79$, $632 = 600 + 30 + 2$ d. $72M+50=100+238$, $117+117=117 \times 2$
	Aspect: Finding the true equation in different circumstances
6	With A replaced by 5, which equation is true? a. $A + 15 = 20$ b. $A - 5 = 20$ c. $20 + A = 15$ d. $15 - A = 20$

Table 4.1 (continued)

Item Number	Test item
	Aspect: Finding the method to solve the equations
7	Which is the method to solve the equation $X + 100 = 100$? a. $(X + 100) - 100 = 100$ b. $(X + 100) \times 100 = 100 \times 100$ c. $(X + 100) - 100 = 100 - 100$ d. $\frac{(X + 100)}{100} = \frac{100}{100}$
	Aspect: Finding the solution to an equation
8	Given $175 = E - 5$, which is the value of E? a. 35 b. 170 c. 180 d. 875
	Aspect: Finding the solution to an equation which relates to the given conditions
9	Which E has the highest value? a. $E - 51 = 28$ b. $E - 31 = 38$ c. $E - 11 = 48$ d. $E - 11 = 58$
10	Which value of X will make $5X = 4X + 9$ more than $5X + 3 = 3X + 9$? a. 3 b. 6 c. 9 d. 12
11	Which equation has different a solution from others? a. $M \div 2 = 23$ b. $9 + M = 55$ c. $4 \times M = 184$ d. $M - 7 = 57$
	Aspect: Selection of an equation which is converted from a verbal problem or a problem which is converted from an equation
12	"A" had X Baht. "B" had 50 Baht more two times of A. The sum of the two equals to four times of A's. What is the equation of the statement? a. $X + 2X + 50 = 4X$ b. $X + 4X - 2X = 2X$ c. $4X - 50 + 2X = X$ d. $X + 2X - 50 = 4X$

Table 4.1 (continued)

Item Number	Test item
	Aspect: Problem solving
13	A man had the cash of 4,650 Baht. After having it deposited in a bank, he had 3,500 Baht remaining. How much money did he deposit in a bank? a. 150 Baht b. 1,100 Baht c. 1,150 Baht d. 8,150 Baht

Measuring Attitude towards Mathematics Computerized Adaptive Testing

To measure attitude towards computerized adaptive testing, a new model for computerized adaptive testing was devised. The model was created using three perspectives of attitude: affective (like or dislike), cognitive (knowledge), and action (Mertens, 1998) with five aspects. The five aspects are: (1) like and interest in CAT, (2) confidence and use of CAT, (3) CAT as modern and useful, (4) CAT as reliable, fair and good, and (5) CAT recommendations. The items relating to each aspect were ordered conceptually by difficulty. They were arranged according to an order of increasing difficulty, as far as possible. This order was tested in Chapter 8. The questionnaire items were grouped under their aspect headings, so that it would be clear to the students what was being asked of them. Thus all items were written in a positive sense, with an ordered response format, from easy to hard. The questionnaire was revised also after discussions with the researcher's supervisor. Ordering is essential for establishing a unidimensional scale as described by Waugh (2002, pp. 67-68).

All items were written in Thai. An English translation that was used for discussions is provided in Table 4.2. The five ordered response categories for the attitude towards mathematics computerized adaptive testing questionnaire for strongly disagree (score 1), for disagree (score 2), for neither agree nor disagree (score 3), for somewhat agree (score 4), and for strongly agree (score 5) were devised to allow consistent and logical discrimination by the respondents.

Table 4.2
Questionnaire on attitude to Computerized Adaptive Testing

Direction: Please read all the following item wordings and answer by making a tick (√) in the box which best describes how strongly you agree or disagree with each wording. For example, if you strongly agree that *Computerized adaptive test is fair for all students*, then √ the strongly agree box. Remember to √ one place for each item.

Item	Item Wording	Strongly disagree	Some-what disagree	Neither agree nor	Some-what agree	Strongly agree
	Aspect: Like and interest in CAT (7 items)					
3	The computerized adaptive testing is very interesting.	☐	☐	☐	☐	☐
14	I am happy doing the computerized adaptive test without limited time.	☐	☐	☐	☐	☐
8	I like the computerized adaptive test because of its immediate feedback.	☐	☐	☐	☐	☐
2	I am happy and enjoyed doing a computerized adaptive test.	☐	☐	☐	☐	☐
12	I feel lucky to have the chance to take a computerized adaptive test.	☐	☐	☐	☐	☐
1	I am enthusiastic about taking part in a computerized adaptive test.	☐	☐	☐	☐	☐
6	I liked the computerized adaptive test because it was not too difficult for me.	☐	☐	☐	☐	☐
	Aspect: Confidence and use of CAT (7 items)					
13	I want a computerized adaptive testing to be used for other subjects.	☐	☐	☐	☐	☐
9	I feel that it is worth having the chance to take the computerized adaptive test.	☐	☐	☐	☐	☐
10	I feel like I am using my full ability with the computerized adaptive test.	☐	☐	☐	☐	☐
11	The computerized adaptive testing makes me want to study Mathematics.	☐	☐	☐	☐	☐

Table 4.2 (continued)

7	After finishing the computerized adaptive testing, I feel like wanting to do another.	☐	☐	☐	☐	☐
5	I believe that I can do the computerized adaptive test well.	☐	☐	☐	☐	☐
4	I took the computerized adaptive test with confidence.	☐	☐	☐	☐	☐
	Aspect: CAT as modern and useful (6 items)					
17	Computerized adaptive testing is modern.	☐	☐	☐	☐	☐
20	The computerized adaptive testing is currently appropriate for these days.	☐	☐	☐	☐	☐
15	Computerized adaptive testing is very useful.	☐	☐	☐	☐	☐
23	Computerized adaptive testing allows students to spend less time on testing.	☐	☐	☐	☐	☐
24	Computerized adaptive testing provides examinees with appropriate items.	☐	☐	☐	☐	☐
18	Computerized adaptive testing saves money.	☐	☐	☐	☐	☐
	Aspect: CAT as reliable, fair and good (5 items)					
21	Computerized adaptive testing is fair for all students.	☐	☐	☐	☐	☐
19	Computerized adaptive testing gives reliable results.	☐	☐	☐	☐	☐
25	Computerized adaptive testing makes examinees careful when doing the test.	☐	☐	☐	☐	☐
16	Computerized adaptive testing is challenging.	☐	☐	☐	☐	☐
22	Computerized adaptive testing inspires the students to do the test.	☐	☐	☐	☐	☐
	Aspect: CAT recommendations (5 items)					
30	I am ready to apply the knowledge from computerized adaptive testing.	☐	☐	☐	☐	☐
28	If possible, I 'd rather take a computerized adaptive test.	☐	☐	☐	☐	☐

Table 4.2 (continued)

29	If I have a chance, I will introduce my younger friends to computerized adaptive testing.	☐	☐	☐	☐	☐
26	I wish I could take a computerized adaptive test in a Mathematics test competition.	☐	☐	☐	☐	☐
27	I will tell my friends about computerized adaptive testing.	☐	☐	☐	☐	☐

Source: Designed by the author specially for this study.

An attitude towards computerized adaptive testing scale was created by analysing the data with the RUMM 2010 computer program. Details of the analysis are discussed in Chapter Eight.

The next chapter (Chapter Five) explains the research methods used in the present research. In Chapter Six the Rasch measurement model analysis results of mathematics items are presented.

CHAPTER 5

RESEARCH METHODOLOGY

This chapter explains the ethics and administrative procedures used. Ethics details administrative approaches including ethics problems. The details of the item bank construction are then explained, as well as Computerized Adaptive Testing (CAT), and attitude measures. Item bank construction concerns the mathematics items, piloting testing, data collection, student samples and data analysis. Computerized Adaptive Testing involves student samples, piloting testing, data collection, and data analysis.

Ethics and Administration

Ethics and Administrative Approvals

The school principals were contacted by the Ubon Ratchathani Rajabhat University President to ask if the schools would be willing to participate in this research. The signed letter of participation, as approved by the Edith Cowan University Ethics Committee, had been returned (Appendix H) indicating a willingness to participate. The letter outlined the purpose of the study and ensured the schools and the students of confidentiality. Students had the right to refuse to participate, and finally, to withdraw from the mathematics testing, the computerized adaptive testing and questionnaire study. Signing the letter indicated satisfaction by the principal, and all participants, of the terms and conditions of the mathematics test computerized adaptive test and questionnaire.

Item 'Bank'

Mathematics Items

Initially, 250 items were used to test an item bank of mathematics on equations. They were separated into six papers with 50 items each. Ten items in each paper were the same (common items), while 40 others were different. The reason why there were six separate papers was that there should be many items in an item bank linked together

on the same scale and the item difficulties should be commensurate with examinees' ability from low to high level.

The subjects in the research were students from elementary (primary) schools, and the items covered a range of difficulty. The first six papers were administered to 2,452 students and analysed with the RUMM 2010 computer program (Andrich et al., 2003) to select items that fit a Rasch model. It was found that insufficient items fitted the measurement model and the researcher constructed one more test with 50 items, of which 40 were new items. Ten others were taken from the items that fitted the measurement model in the previous test. The seventh paper was administered to 610 students, and the RUMM 2010 computer program was used to test for fit to the measurement model. Analysis was made to find out the quality of the test and also fit to the model. The items fitting the model were stored in the item bank to be used for the computer adaptive testing.

The structure of the test was categorised according to the objectives and the time used as shown in Table 5.1.The researcher constructed the mathematical tests on the equations, based on the following aspects.

1. Objectives, contents, and time required in learning and teaching mathematics (equations) were studied from the curriculum and the handbook prescribed for the Year 6 students.

2. There was a table of the items according to the objectives, starting from the basic to the complex in accordance with the teaching periods (Table 5.1).

3. The multiple-choice items were constructed with four alternatives, but only one correct answer. The correct answer gets one point, and a wrong one gets zero.

4. The mathematics test items for the item bank consisted of 290 items constructed according to the sub-objectives in Table 5.1. The 290 items were divided into seven papers with 50 items each, and each paper contained 40 different items and 10 common items in which the researcher utilized, as suggested by Lord (1980), that the common items should be 20 percent of the whole test items; thus, 20 percent of 50 items is 10 items. For the seventh paper, ten items fitting the model were selected from the first six papers and then Rasch analyzed. All the items in the seven tests were arranged from easy to hard.

Table 5.1
The structure of the mathematical test as catagorised according to the sub-objectives and the time used.

Objectives	Time used in teaching (periods)	Number of items
1. Given several symbolic sentences, students can identify the equation. 2. Given several equations, students can identify the true equation	2	3
3. Given several equations, students can identify the equation with unknown identify.	2	3
4. Given an equation with unknown identity, students can choose the number and substitute the unknown identity.	2	3
5. Given an equation with the unknown identity on addition, subtraction, multiplication and division, students can tell how to find the solution, and solve the equation correctly.	17	25
6. Given a verbal problem of daily life which require addition, subtraction, multiplication or division, students can convert the verbal problems into an equation, solve it and get the answer.	11	16
Total	34	50

Piloting Testing for the Item Bank

The mathematics test items were checked several times by the researcher until a satisfactory test covering all objectives were included. There were two steps for pilot testing of the test. For the first step, the items were checked by the experts in assessment and mathematical content in regard to conformity between the test items and the behavioural objectives of the syllabus, in order to ascertain that the constructed test

items accorded with the objectives which means that the test has sound content validity. In addition to validating the content, the experts made a few adjustments in the language use in problem solving test items. For example, 'cm.' the contracted form was suggested to be change into the full word as centimetre. Some items were changed to make the meaning more clear, for example, from "Adam bought Z pieces of pork, costing 3 Baht each. *The sum used was 54 Baht.* Which equation show how many pieces did he buy?" became "Adam bought Z pieces of pork, costing 3 Baht each. *He spent 54 Baht for all.* Which equation showed how many pieces did he buy?"The test items were then corrected accordingly to be more accurate and satisfactory for Prathom Suksa (Year) 6 students.

For the second step, the test was then taken by 35 volunteer Prathom Suksa 6 students covering those who were the smartest, average, and poor in mathematics. The students answered the seven test papers, five students per each paper. It was found that the students could do every test within the averaged time, 40-60 minutes and handed in their answers to the researcher, and then the researcher asked them whether they had any difficulties in understanding questions or in responding to what they expected and what they actually did in each question. They all replied that they understood and cleared all but some items which were hard for them, particularly the complicated ones. For example, the following two items were queried by the students.

1. Which choice has two equations with the same solutions?
   a. $3X - 7 = 23$, $2X - 10 = 8$
   b. $4X + 3 = 31$, $5X - 3 = 32$
   c. $7X + 2 = 58$, $9X - 4 = 41$
   d. $4X - 7 = 21$, $3X + 6 = 33$

2. A teacher wants to divide 100 boy scouts into equal groups with 9 members a group, while keeping one scout free from any group. How many groups will he use?
   a. 10 groups
   b. 11 groups
   c. 12 groups
   d. 13 groups

Student Samples for the Item Bank

The subjects in the study were 3,062, Prathom Suksa 6 students from 85 classes of 21 public schools in Ubon Ratchathani Province. There were 2,452 students who took part in the first sampling (test 1 to test 6) and 610 students participated in the

second sampling (test 7). The subjects were derived through stratified random sampling as given in Table 5.2.

Table 5.2
Samples by school for item bank tcesting

Schools	Number of students	Percentage
1	387	12.64
2	365	11.92
3	351	11.46
4	319	10.42
5	300	9.80
6	150	4.90
7	129	4.21
8	121	3.95
9	120	3.92
10	115	3.76
11	109	3.56
12	90	2.94
13	87	2.84
14	78	2.55
15	73	2.38
16	66	2.16
17	51	1.67
18	44	1.44
19	39	1.27
20	38	1.24
21	30	0.97
Total	3,062	100

Data Collection for Item Bank

In terms of data collection, the researcher presented the letter of cooperation to the director of the schools whose voluntary students were the sample subjects. Parents of the students were asked to allow their children to answer the test. Coordination was made with the schools to clarify the research objectives and set dates and time to test the students. The total of 2,452 students from the Prathom Suksa 6 were tested with the first six papers. There were 409, 413, 412, 400, 410, and 408 students who took part in the 1st to the 6th tests respectively.

For the test administration, the researcher handed out the mixed six test papers to the students in each class. Answer sheets and attached papers were given to the students with the explanation about the time allocated and how to answer the equations to make sure that they understood the directions of each section. There were no problems from the students because most of them were familiar with this kind of testing. The students were informed about the time allocated twice, one after half an hour of the testing and the other, five minutes before the finished time. Five minutes before the finished time, the students could thoroughly examine the test papers again before submission. The same procedure was repeated in the second testing with the 7th paper with 610 Prathom Suksa 6 students.

Test Marking

In relation to scoring, the researcher made the key answers for the seven tests papers and rechecked the correctness of the answers. The scoring was done carefully in order to prevent mistakes and to guarantee the accuracy. The researcher's colleagues were asked to score and check the whole process of scoring. The result turned out satisfactorily because the objectivity of the papers was guaranteed by having no mistakes in scoring from each of the scorers.

Data Entry for the Item Bank

Responses for the mathematics tests from 2,452 Prathom Suksa 6 (Grade 6) students were entered into the Excel program, as per the response category codes (zero for wrong, one for right, and nine for missing) and then converted to a text file. This was checked twice to ensure accuracy. For the mathematics test, there was a student

code number and there were the item numbers in the same row from left to right. The student code number started from 1001 to 3452. The data pattern had 254 columns: columns 1-4 were for the student code number; columns 5-14 were for 10 answers of common test items; columns 15-54 were for 40 answers of test 1; columns 55-94 were for 40 answers of test 2; columns 95-134 were for 40 answers of test 3; columns 135-174 were for 40 answers of test 4; columns 175-214 were for 40 answers of test 5; and columns 215-254 were for 40 answers of test 6. For the students who took part in test 1, the data for column 55-254 were missing data (9s). The item layout, as provided in the Excel program, is shown in Table 5.3.

Table 5.3
Mathematic data sample (Excel program) for the 1st to 6th tests

Column number							
1-4	**5-14**	**15-54**	**55-94**	**95-134**	**135-174**	**175-214**	**215-254**
Student Number	10 common item	40 answers of test 1	40 answers of test 2	40 answers of test 3	40 answers of test 4	40 answers of test 5	40 answers of test 6
1001-1409	101..000	101..010	999..999	999..999	999..999	999..999	999..999
1410-1822	100..111	999..999	101..011	999..999	999..999	999..999	999..999
1823-2234	101..000	999..999	999..999	101..010	999..999	999..999	999..999
2235-2644	001..111	999..999	999..999	999..999	101..000	999..999	999..999
2645-3044	111..010	999..999	999..999	999..999	999..999	001..101	999..999
3045-3452	000..101	999..999	999..999	999..999	999..999	999..999	100..111

In order to examine the correctness of data entered into the Excel file, the researcher checked them again. There were a few mistakes in the students' answer and they were corrected before the next process was employed. After all data were entered into the Excel files, it was converted into a word text document ready for Rasch analysis, described in Chapter Six.

Data Analysis for Item Bank

The data were analysed with computer program RUMM 2010. The RUMM 2010 computer program (Andrich et al., 2003) is currently the best of the main computer programs for Rasch measures. The RUMM 2010 program provides a comprehensive set of output data to test many aspects of the conceptual model of the variable, the answering consistency of the response categories, both item and person fit

to the measurement model, and targeting as presented in Chapter Six. The researcher then checked whether the test items fitted the measurement model. It was found that 172 out of 250 items did not perform according to the measurement model. Therefore, they were deleted from the scale, leaving 78 items that fitted the measurement model.

Only 78 items were initially stored in the item bank. More items were needed for the item bank to be effective when used with the Computerized Adaptive Test. Hence, the researcher created another 40 mathematics items. To access the item bank efficiently, these two sets of the items were linked. For linking the scales, the researcher chose 10 items from the set of 78 items. The 10 items and the 40 items were then tried out with the group of 610 students mentioned above. The data from these 50 items from the 610 students were entered into an Excel program (as shown in Table 5.4), as per the response category codes (zero for wrong, one for right, and nine for missing) and then converted to a text file for analyses with RUMM 2010. Of these 50 items, 30 were deleted as not fitting the Rasch measurement model, leaving 20 good fitting items to be added to the set of 78 items (the results as shown in Chapter Six). The good-fitting 98 items were stored in the item bank by using the software that the researcher created.

Table 5.4
Mathematics data sample (Excel program) for the 7th test

Item no. / Id number	1	2	3	4	5	...	50
1001	0	1	1	0	1	...	
1002	1	0	0	0	1	...	
:						...	
:						...	
1610						...	

Setting Up the Item Bank

After obtaining 98 test items that fitted the measurement model, the researcher installed them in the item bank. The detailed test items included item number, stem and choices, difficulty, standard error (SE), and key answer (1 for choice a, 2 for choice b, 3 for choice c, and 4 for choice d) for each item. In order to be certain that all of 98 items were correct, as in their original versions, the researcher copied each of the items and

saved them as a graphic file, each file consisted of one item, that is, 98 files for 98 items.

After all these processes, the researcher copied each item and pasted it into the area of the bank prepared for typing of each test item in order to prevent the mistakes derived from typing. Details such as item number, difficulty and key answers were added until 98 items were completed. The examples before and after storing an item in the bank are shown in Figures 5.1-5.2.

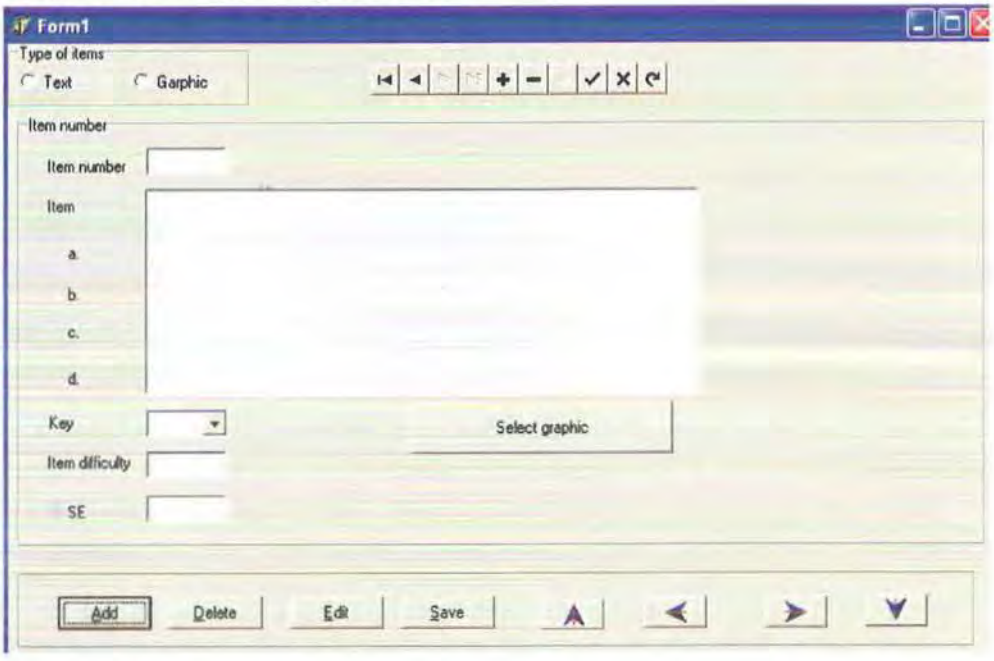


Figure 5.1 **A blank form before storing an item in the bank**

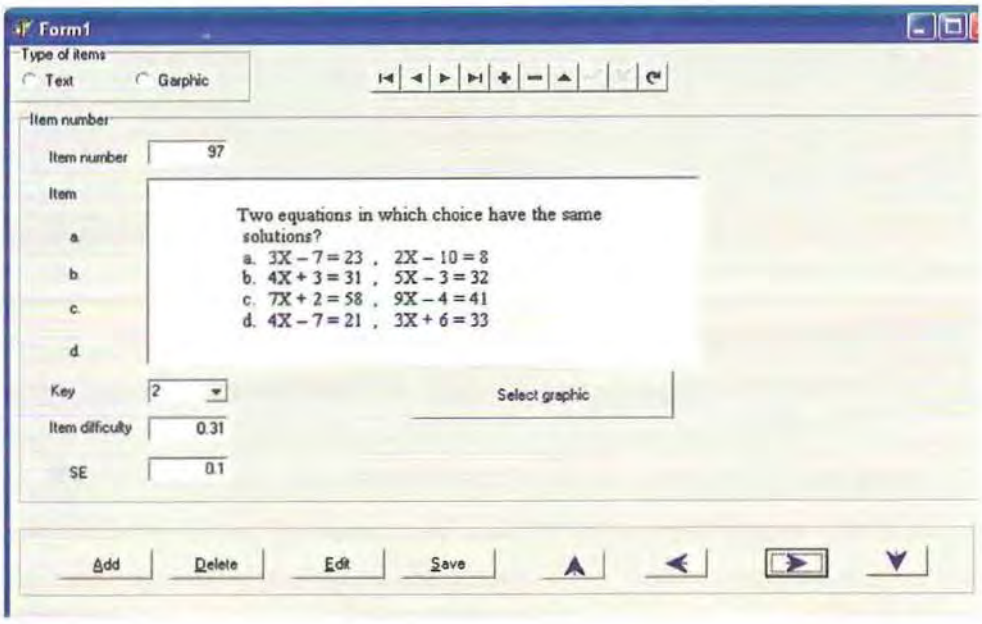


Figure 5.2 **The completed form of item 97 after storing in the bank**

Student Samples for the CAT Testing

Subjects for testing CAT were 400 gender-mixed Prathom Suksa 6 students from two public schools in Ubon Ratchathani Province, in which a computer laboratory was provided. The subjects gained from a stratified random sampling technique, using students' mathematics competency. According to the Mathematics National test scores, the students were divided into three groups of good, fair and poor competency. The good students were those who got 75 % or more, the fair were between 50-74% and the poor were 50% and less. Table 5.5 shows the details of the students by competency group and school for testing the Computerized Adaptive Test program.

Table 5.5
Number of students by competency and school for testing CAT

Competency Group	Number of students			Percentage		
	School 1	School 2	Total	School 1	School 2	Total
good	70	50	120	17.50	12.50	30
fair	90	70	160	22.50	17.50	40
poor	70	50	120	17.50	12.50	30
Total	230	170	400	57.50	42.50	100

Construction of the Computerized Adaptive Test

To create the computer program for the Mathematics Computerized Adaptive Testing, the researcher followed the following three steps.

Step 1 Selection of a computer language in writing the programme. The Delphi language was selected because it is one of high performance computer languages, having high abilities in database management and scientific data estimation.

Step 2 Creation of a flowchart for the Computerized Adaptive Testing administration which covered two parts. Part I was a programme registration. It

provided different functions shown in Figure 5.3a. Part 2 was a test administration which contained different functions shown in Figure 5.3b.

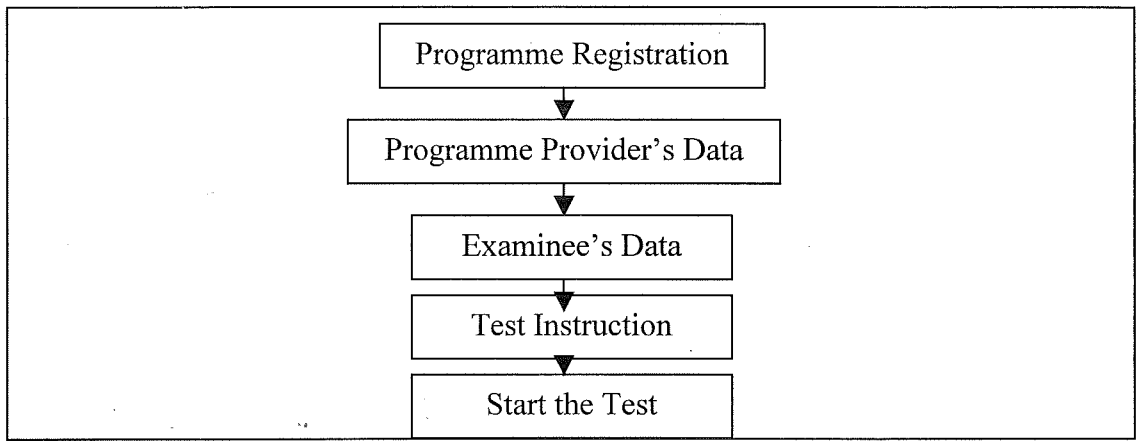


Figure 5.3a **Part I of the CAT program registration**

Source: Devised by the researcher for this study.

The Computerized Adaptive Testing program registration part, shown in Figure 5.3a, included such functions as the program registration, program provider's data, examinee's data and test instruction. The program provider's data displayed name of the programme, in this case, The Chaow Computerized Adaptive Test Computer Program as well as name of the researcher and the university. The examinee's data required a student's information, i.e., student code, name, grade, school and password (security system).

In the test instruction (see Appendix I), details of the Computerized Adaptive Testing program and how to accomplish the test were explained. Examples of the Computerized Adaptive Testing program registration were also displayed in Appendix I. Then the examinee clicked the "next" button to start the test. The administration of the test is described in Figure 5.3 b.

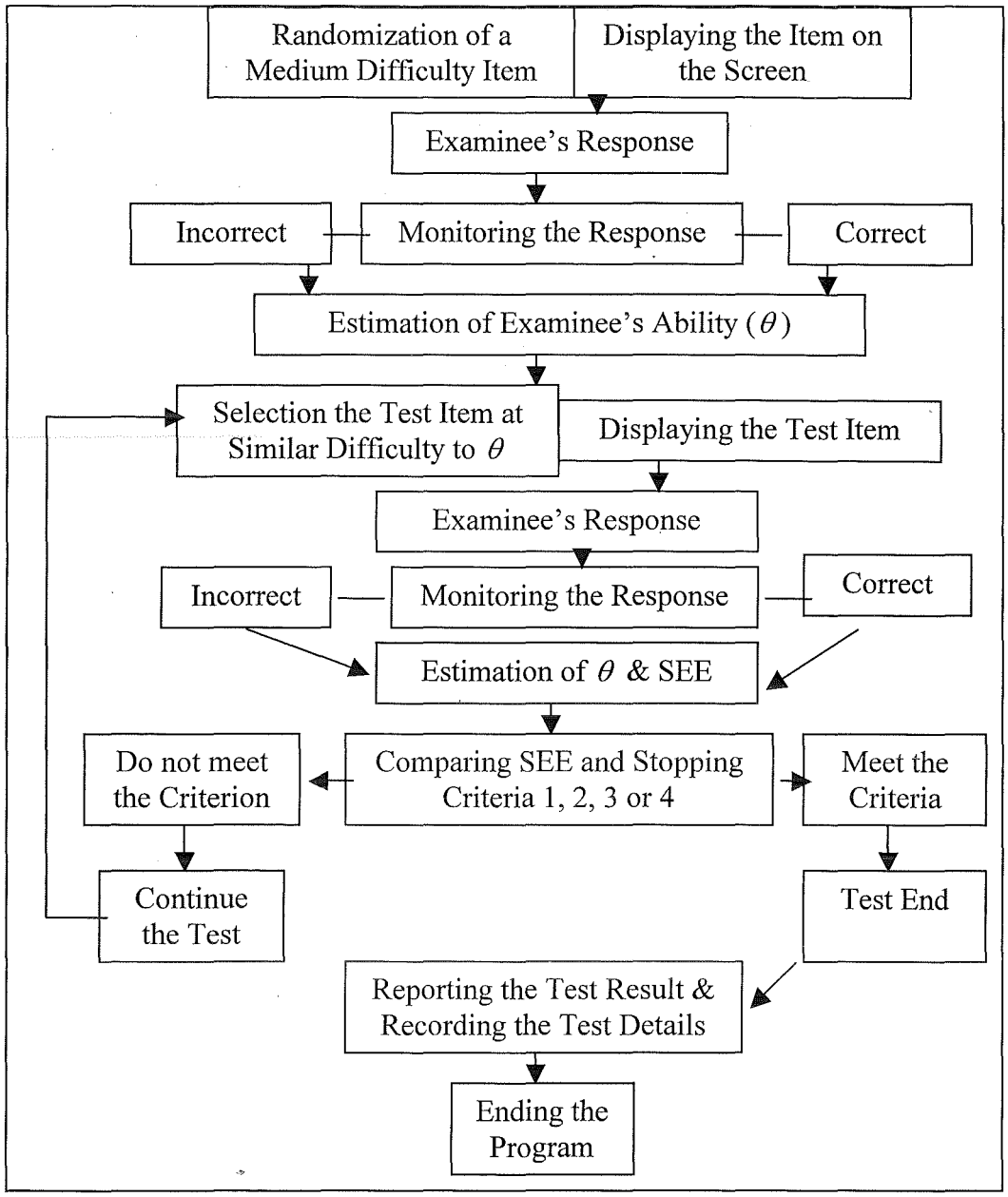


Figure 5.3b **Part II of the construction of the CAT program test**

Source: Adapted from Maneelek (1997) and Supeesut (1999).
Note: SEE = Standard Error of Estimation.

The activation of part II of the Computerized Adaptive Testing program, as seen in Figure 5.3b, started with the initial item which was randomised as a medium difficulty item displayed. The choices were provided at the bottom. The examinee's response would be monitored. The correct or incorrect response would be a source for the estimation of the examinee's competency by using Bayesian Updating (Owen, 1969, 1975). The equations are detailed in Chapter Three Theoretical Framework.

According to the ability estimated (θ), the Computerized Adaptive Testing program would select and display an item with the difficulty close to the examinee's

estimated competency. Based on this response, the examinee's ability (θ) and the Standard Error of Estimation (SEE) would be calculated. The Standard Error of Estimation would be compared with the stopping criteria 1, 2, 3 or 4. If the score did not meet the criteria, another test item would pop up and repeat the monitoring and the estimation functions. Such interaction would keep going until the Standard Error of Estimation met the criteria. After the test ended, the test result and details would be automatically saved in the computer and reported on the screen.

Step 3 Writing up the program based on the flowchart created in Step 2.

Pilot Testing of CAT

The process of pilot testing of the Computerized Adaptive Testing covered three stages as follows.

Stage 1. Examining the accuracy of the test results gained from program computing and manual calculation of the fresh data that was done several times until the researcher was certain of its accuracy.

Stage 2. The Computerized Adaptive Testing program was piloted with four students of Prathom Suksa 6 in Ubon Ratchathani Province by Computerized Adaptive Testing using four different stopping criteria. After the pilot test, the researcher asked the students whether they had any problems in taking the test. There had been a problem with the first item of the test due to their excitement because they had never done this before. This caused them to take a longer time in answering the item. In addition, some items were not readable on the computer screen especially those with long problems or complicated equations. These problems were solved after the size of the alphabets were made smaller and recorded into the item bank.

Stage 3. The Computerized Adaptive Testing program was experimented with twelve students from Prathom Suksa 6 in Ubon Ratchathani Province. These students were of mixed competency: high, medium, and low levels. Each of four different stopping criteria was applied to the three students. The students spent an average time of 20 minutes in taking the test with the test items ranging between three to nine items. The experiment with the test takers showed that Computerized Adaptive Testing was very exciting due to its newness. However, some students suggested that the questions should have been easier. Also, they wanted to have more computers in their schools. In

terms of the clarification of the test items and instructions, the students stated that they were clear and easy to follow.

Administration of CAT

The process of Computerized Adaptive Testing administration started from getting the results of a sample of the national test scores on mathematics from two sampled schools in order to classify the students into three different groups of mathematical competency, high, medium, and low.

A total of 665 students, based on the scores gained from the national test, were obtained. The researcher then selected 400 students and divided them into four sub-groups having the same features of mathematical competency as high, medium, and low levels by using a stratified random sampling technique. Each group contained 100 members whose different mathematical competencies were mixed: 30 students in high ability group, 40 in medium, and 30 in low level. Four stopping criteria techniques were used for simple random selection of the students for each group. Each student's information such as name and surname, student code, class, and school was installed into the database of the Computerized Adaptive Testing program with passwords by the researcher. Then the students' passwords were told to them before they entered the testing. The researcher set the computers up for Computerized Adaptive Testing by installing the Computerized Adaptive Testing program into the computers on the basis of four stopping criteria. After this process, each of the students was allowed to enter the room and sit in front of the computers in accordance with the specific computer for him or her. After that, the researcher explained Computerized Adaptive Testing and how to do the test, and demonstrated answering the test through the projector. They were allowed to ask questions about the testing. The students were then told to begin the test by following the instructions mentioned on Computerized Adaptive Testing from the beginning until the last process. After the completion of the test, the result of each of the test takers was automatically recorded into the computer for further statistical analysis. Examples of the Computerized Adaptive Testing administration are presented in Appendix I.

Data Analysis of CAT

After the students finished the test, the researcher coded and keyed in necessary information of 400 students' data as required in the SPSS program and kept them in a data file. Details of student's data, such as student number (001-400), and important variable data such as testing time, test length, number of answered items, examinee ability, and types of stopping criteria were included. After the researcher keyed in all the data, it was checked for accuracy. The next process, data analysis was then implemented.

The SPSS computer program (Pallant, 2001) was used to analyse data from the 400 Prathom Suksa 6 students. The frequencies and percentages of mathematics ability were used as the indicators to examine mathematics competencies of the students. A One Way ANOVA was used to examine differences in test length and testing times among the different groups relating to stopping criteria and mathematics competencies, and also to examine any differences mathematics competencies among the different groups for stopping criteria. ANOVA is the appropriate statistic to use because there are more than two groups of the students and test length, testing times and mathematics competencies were measured on ratio or interval scales (Cavana, Delahaye, & Sekaran, 2001). Because the F statistics were significantly different, the Sheffe Multiple Range test was used to determine between which groups the true differences lie (Cavana et al., 2001). Details of the results of the data analysis are presented in Chapter Seven.

Attitude to Computerized Adaptive Testing

Pilot Testing of Questionnaire: Attitude towards CAT

An informal trial of the questionnaire was conducted with five students. They were asked to answer the questionnaire, and then the researcher discussed the questionnaire with them. Their feedback indicated that the questionnaire would have been found better if it had been made in a regular type of letters instead of italic. Italicized letters made them difficult to be read. In addition to the type of the letters, a tick should have been allowed to be put into the space provided in the questions instead of into the box (☐). They stated that the instructions for item wordings were clear enough and that Prathom Suksa 6 students should be able to understand the items and answer them satisfactorily. All the weaknesses were improved as suggested by the

students before the questionnaire was implemented. The questionnaire was then considered ready for a formal pilot test.

A formal pilot test of the attitude questionnaire was conducted with 12 students of Prathom Suksa 6 based on certain conditions: (a) the respondent can read and understand the questions or items; (b) the respondent possesses the information to answer the questions or items; (c) the respondent is willing to answer the questions or items honestly (Wolf, 1997). The researcher selected voluntary students and explained to them what Computerized Adaptive Testing is and its processes. After that, the students were asked to try the mathematics Computerized Adaptive Testing. The questionnaire was first explained to the students by the researcher showing how to reply to each response category, and then 40 minutes was given to the students to complete the questionnaire. It was found that all students could complete their questionnaire, ranging from 20 to 30 minutes and they handed in the results to the researcher. The researcher then asked them whether they had any difficulties in understanding wordings or in responding to what they expected and what they actually did in each question. They all replied that they clearly understood everything. Moreover, they said that they liked taking Computerized Adaptive Testing and considered it newer, more fun, exciting, and challenging. They would like to take Computerized Adaptive Testing with other areas of study. Therefore, the researcher did not discard any items. Students made no additional comments about the questionnaire in general, no comments were made that any important aspects had been left out, and no other main comments were made about the questionnaire. The response format was satisfactory and the instructions were understandable.

Student Sample for Attitude towards CAT

The 400 students who responded to the questionnaire, Attitude towards Computerized Adaptive Testing, were in Prathom Suksa 6 from schools in Ubon Ratchathani Province. The schools were purposively selected because of their readiness in computer operational rooms. The sample was the same group as those who sat mathematics Computerized Adaptive Testing as previously mentioned shown in Table 5.5.

Data Collection for Attitude towards CAT

In terms of data collection, the researcher presented the letter of cooperation to the director of the schools whose voluntary students were the sample subjects. Parents of the students were asked to allow their children to answer both the mathematics Computerized Adaptive Testing and the attitude questionnaire. Coordination was made with the schools to clarify the research objectives and set dates and time to test the students. The total of 400 students from the Prathom Suksa 6 students answered the attitude questionnaire. After the students finished taking the mathematics Computerized Adaptive Testing, the attitudinal survey questionnaire was distributed to them. The researcher explained to them how to answer the questions in the questionnaire. There were no problems and difficulties in these processes.

Data Preparation for Attitude towards CAT

After the administration of the questionnaire to the convenience sample of 400 Prathom Suksa 6 students, the checked questionnaires were collated by student code number and entered into an Excel program. For the attitude questionnaire, there was the student code number and there were the item numbers in the same row from left to right. The student code number started from number 1001 and continued until 1400 (400 students). The item numbers started from 1 and continued until 30 (30 items). Item responses ranged from 1 to 5. The data were entered into an Excel program as shown in Table 5.6.

Table 5.6
Attitude questionnaire data sample (Excel program)

Item no. / **Student code**	**1**	**2**	**3**	**4**	**... 30**
1001	2	3	2	3	
1002	3	1	3	2	
1003	2	3	1	2	
:					
:					
1400					

Student responses were entered 1 (for strongly disagree), 2 (for disagree), 3 (for neither agree nor disagree), 4 (for somewhat agree), and 5 (for strongly agree) and 9 (for missing). To interpret these numbers, the first student, 1001, chose 2, 3, 2, 3 for items 1, 2, 3, 4, respectively, meaning that he/she 'disagreed' (2), 'neither agreed or disagreed' (3), 'disagreed' (2), and 'neither agreed nor disagreed' (3) with items 1, 2, 3 and 4, respectively.

In order to examine the correctness of data collected, the researcher checked them thoroughly again and they were corrected before the next process was employed.

After all data were entered into an Excel file, it was converted into a word text document ready for Rasch analysis, described in Chapter Eight.

Data Analysis for Attitude towards CAT

After rechecking the collected data, the researcher analysed them with program RUMM 2010. The RUMM 2010 computer program (Andrich et al., 2003) is currently the best of the main computer programs for Rasch measures. The RUMM 2010 program provides a comprehensive set of output data to test many aspects of the conceptual model of the variable, the answering consistency of the response categories, both item and person fit to the measurement model, and targeting as presented in Chapter Eight. It also provides some wonderful graphical output for various aspects such as Item Characteristic Curves, Response Category Curves and person measure-item difficulty map.

The next chapter (Chapter Six) describes the data analysis for the mathematics item bank using the Rasch Unidimensional Measurement Model (RUMM 2010) computer program.

CHAPTER 6

DATA ANALYSIS (PART I)
MATHEMATICS ITEM BANK

This chapter describes the process of data analysis for the mathematics item bank, using the Rasch Unidimensional Measurement Model (RUMM) computer program (Andrich et al., 2003). The initial analysis with 250 items comes from six tests with 50 items each. For linking the scales, each test contained 10 common items first, and then the six data sets were combined. Responses for the mathematics tests from 2,452 Prathom Suksa 6 (Grade 6) students were entered into an Excel file, as per the response category codes (zero for wrong and one for right) and then converted to a text file. The data pattern had 254 columns: columns 1-4 were for the ID; columns 5-14 were for 10 answers of common test items; columns 15-54 were for 40 answers of test 1; columns 55-94 were for 40 answers of test 2; columns 95-134 were for 40 answers of test 3; columns 135-174 were for 40 answers of test 4; columns 175-214 were for 40 answers of test 5; and columns 215-254 were for 40 answers of test 6. The non-performing items of the mathematics test (172 items out of 250) were deleted from the scale, leaving 78 items that fitted the measurement model.

Because the 172 items (out of 250) were deleted, as not fitting a Rasch measurement model, only 78 items were stored in the item bank. To make the Computerized Adaptive Test, a computer program designed by the researcher to help Prathom Suksa 6 students access the item bank efficiently, more items were needed for the item bank. So the researcher created more 40 mathematics items. For linking the scales, the researcher included common items by choosing 10 items from the 78 set. The 10 items were added to the 40 set for calibration together and linking with 78 items. The data from 610 students were analysed using the RUMM computer program. Of these 50 items, 30 were deleted as not fitting a Rasch measurement model, leaving 20 good fitting items to be add to the set of 78 items.

The presentation begins with two descriptions of the analysis for the mathematics achievements that are reported for 78 items in the first and 20 items in the

second. The Rasch analysis provides data on global item and person fit to the measurement model, individual item fit, dimensionality, reliability, Student Separation Index and targeting. A summary list of the main findings is presented at the end of the chapter.

In Rasch analysis, the items are designed in a conceptual order by difficulty and this order is tested. The data for the items have to also fit the measurement model in order to create a linear scale and this is tested. The person measures and item difficulties were calibrated on the same scale by the RUMM 2010 program, thus providing the creation of a linear measure of mathematics achievements.

The results of the analysis are set out in Tables 6.1 and 6.2, and Figures 6.1 to 6.4. Table 6.1 presents a summary of the global fit statistics of the measure of mathematics achievement in the first and second testings, including the item-trait test of fit to the measurement model. The item difficulties in order of the 78 items are shown in Table 6.2. Figure 6.1 shows person measures of ability and item difficulty map for the mathematics test (78 items, 2,452 students), with the mathematics measures on the LHS and the difficulties on the RHS. Figures 6.2 and 6.3 show response category curves for item 76 (good-fitting item) and item 180 (not-so-good fitting item). Figure 6.4 shows item locations on the lower side (LS) and mathematics measures on the upper side (US) on the same scale in logits for the test. Appendix C shows, in probability order, the location on the continuum, fit to the measurement model and probability of fit to the model for the 78 items.

Rasch Analysis: 78 Items Scale

Global Fit to the Measurement Model

The final analysis with the RUMM program tested the 78 items (N=2,452) in order to create a linear scale of mathematics achievement from an initial bank of 250 items. The residuals were examined; the residuals being the difference between the expected item score calculated according to the Rasch measurement model and the actual item score of the students. This is converted to a standardized residual score in the computer program. The global item fit residuals and global student fit residuals have a mean near zero and a standard deviation near one, when the data fit the measurement

model. In this case, the global item and person fit residuals indicate a satisfactory, but not excellent, fit to the measurement model (see Table 6.1).

Individual Item Fit

The individual probability of fit of items to the measurement model was then checked to identify items that fitted the model (see Appendix C). Of the 78 items, 71 fitted the measurement model with probability $p>0.04$.

Item Trait Test-of-Fit

The item-trait test of fit examines the consistency of the item difficulties across the student mathematics measures along the scale. This determines whether there was agreement among students as to the difficulties of all items along the scale. The item-trait interaction was not statistically significant at 0.01 level [Chi-square (df =690) =760.34, p =0.03]. This means that a dominant trait was measured and that overall fit to the measurement is acceptable, but not excellent.

Targeting

The item difficulties range from –1.3 logits (SE=0.12) to + 1.6 logits (SE=0.14) and the student measures range from –3.4 logits to +4.2 logits. There are some students (34%) whose mathematics abilities are more than +1.6 logits and less than –1.3 logits and hence not 'matched' against an item location on the scale.

In Figure 6.1, there are no items matching persons at either the lowest end (-1.5 to –3.5 logits) or the highest end (+1.5 to +4.4 logits) of the scale, indicating the improvements that are needed for the test. That is, both easy items and hard items need to be added to improve the targeting of the item for these Prathom Suksa 6 students.

Table 6.1
Summary of fit statistics for mathematics achievement scale (78 items)

	Items	students
Number	78	2,452
Location mean	0.00	0.58
Standard deviation	0.62	1.64
Fit statistic mean	0.63	0.08
Fit statistic standard deviation	1.23	0.73
Item-trait interaction chi square = 760.34		
Degrees of freedom = 690		
Probability of item-trait (p) = 0.03		
Student Separation Index = 0.83		
Power of test-of fit: Good (based on the Separation Index)		

Notes on Table 6.1

1. The item means are constrained to zero by the measurement model.
2. When the data fit the model, the fit statistics approximate a distribution with a mean near zero and a standard deviation near one. The item fit and student fit are satisfactory, but neither is an excellent fit.
3. The item-trait interaction indicates the agreement displayed with all the items across all students from different locations on the scale (acceptable for these data). This means that a dominant trait has been measured.
4. The Student Separation Index is the proportion of observed student mathematics variance considered true (in this scale, 83% and is acceptable). It tells us that the measures are well separated compared to the errors.
5. Numbers are given to two decimal places because the errors are between 0.11 and 0.14.

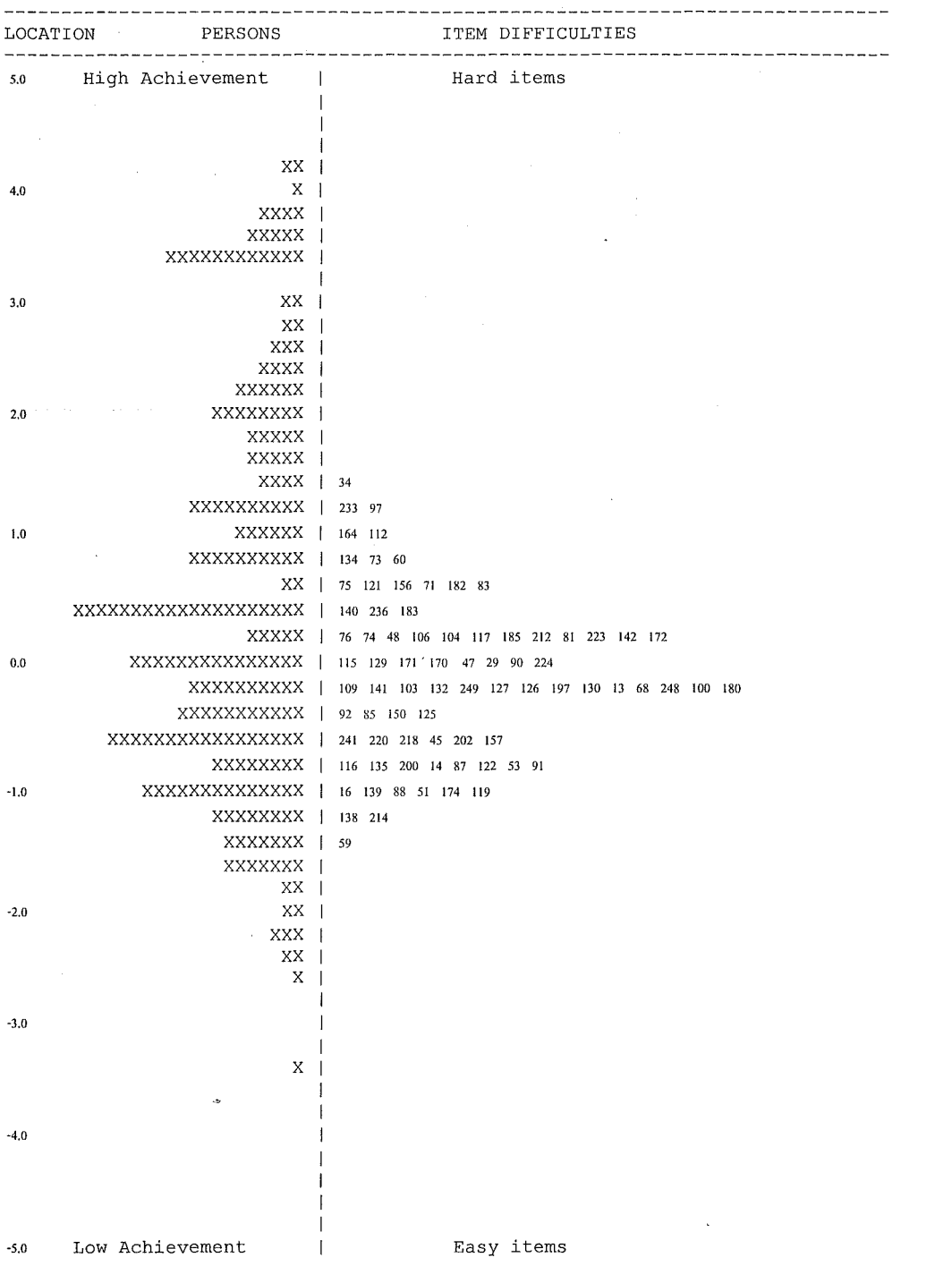


Figure 6.1 **Person measures of achievement and item difficulty map for mathematics test (N=2,452, I=78)**

Notes on Figure 6.1

1. The scale is in logits, the log odds of answering positively.
2. Mathematics measures are calibrated on the same scale as the item difficulties.
3. Measures are ordered from low to high on the LHS and item difficulties are ordered from easy to hard on the RHS.
4. Items at the easy end of the scale are answered positively by most students. As the items become harder, students need a higher mathematics measure to answer the items positively.
5. Each x represents 11 students.

Category Response Curves

The RUMM program provides a category response curve for each item, which makes it possible to view the ordering of the thresholds, and to check whether the category responses are being answered consistently and logically. A perusal of the category response curves for the 78 items indicates that the students answered the response categories consistently and logically. The items contained two response categories earning 0 and 1 mark. Figure 6.2 shows the category response curve for the good fitting item 76. The category 0 curve refers to a 0 mark (category response wrong) and the category 1 curve refers to 1 mark (category response right).

Item 76 is a good-fitting item with a chi-square probability of 0.94. Its difficulty is +0.21 and this means that the students found that the item is relatively hard for them. Figure 6.2 shows that the category curve 0 (category response wrong) indicates that when a student has a very low mathematics measure (-4 logits), then the probability of answering in this category (getting 0) is 0.98 (very high as expected). As the student's mathematics measure increases to about -1 logits, then the probability of scoring 0 drops to near 0.80 (as expected). If the student's mathematics measure increases to about 0 logits, then the probability of scoring 0 drops to near 0.50 (as expected). When the student's mathematics measure increases to about +4 logits, then the probability of scoring a 0 mark drops to zero (as expected).

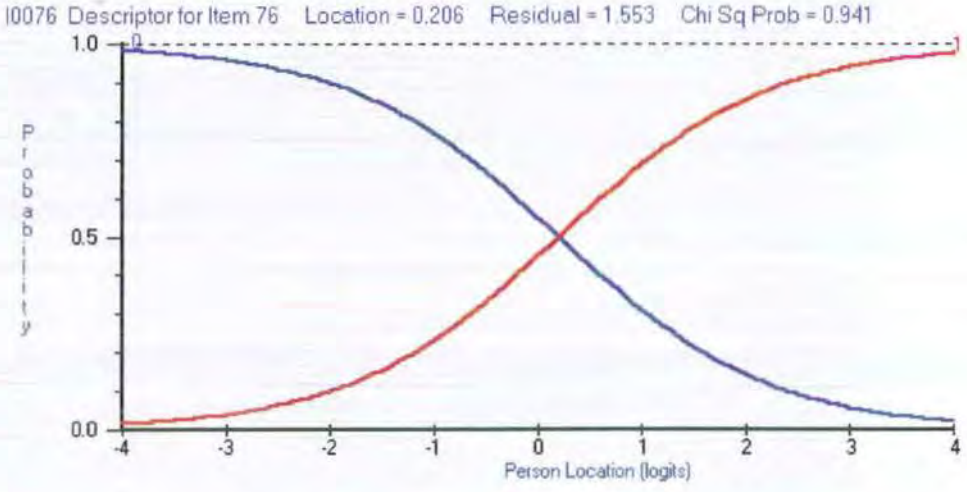


Figure 6.2 **Response category curve for item 76 (good-fitting item)**

For curve 1 (category response right), when the student has a very low mathematics measure (-4 logits), then the probability of answering right (getting 1) is near zero (very low as expected). When the student mathematics measure increases to -2 logits, then the probability of answering right (getting 1) increases to 0.1 (as expected). When the student mathematics measure increases to 0 logits, then the probability of answering right increases to near 0.5 (as expected). When the student mathematics measure increases to +4, the probability of answering right increases to 1 (as expected).

Item 180 is a moderately difficulty item (difficulty = -0.01 logits) that doesn't fit the measurement model as well as one would like. Nevertheless, the Response Category Curve is good. Figure 6.3 shows that the category curve 0 (category response wrong) indicates that when students have very low mathematics measure (-4 logits), then the probability of answering in this category (getting 0) is 0.98 (very high as expected). As the student's mathematics measure increases (to -1 logits), then the probability of getting 0 drops to near 0.75 (as expected). If the student's mathematics measure increases (to 0 logits), then the probability of getting 0 drops to near 0.52 (as expected).When the student's mathematics measure increases to+4 logits, then the probability of getting 0 drops to zero (as expected).

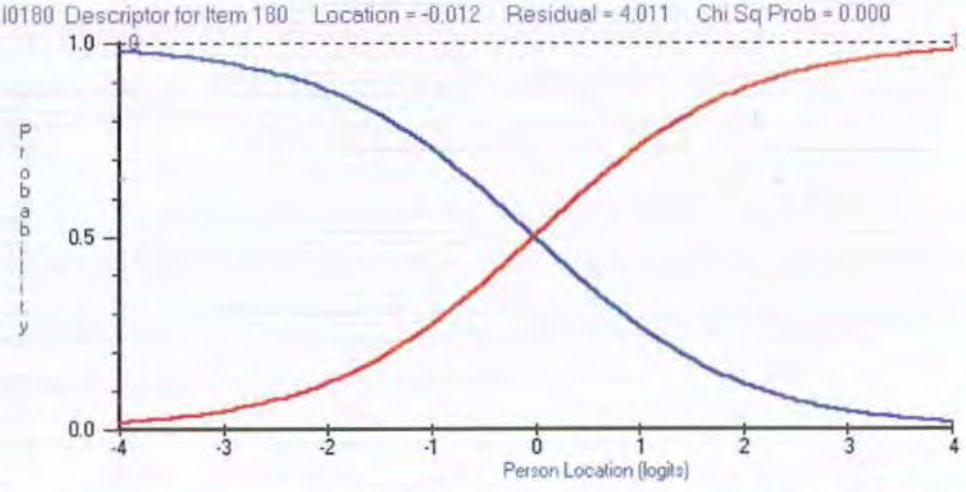


Figure 6.3 **Response category curve for item 180 (not-so-good fitting item)**

For curve 1 (category response right), when the student has a very low mathematics measure (-4 logits), then the probability of answering right (getting 1) is near zero (very low as expected). When the student mathematics measure increases to -2 logits, then the probability of answering right (getting 1) increases to 0.12 (as expected). When the student mathematics measure increases to 0.0 logits, then the probability of

answering right increases near 0.5 (as expected). When the student mathematics measure increases to +4, the probability of answering right increases to 1 (very high as expected).

Person Measure/ Item Difficulty Scale

The linear scale of mathematics achievement (Figure 6.4) shows the student measures on the top side from a low of -4.0 logits (left hand side) to a high of +4.4 logits (right hand side). The item difficulties are calculated on the same scale as the student measures on the bottom side from easy (-1.4 logits) to hard (+1.7 logits). There are approximately 600 students who found these test items easy and approximately 180 who found who found them hard. The item difficulties were appropriate for the rest of the students, approximately 1,770 students.

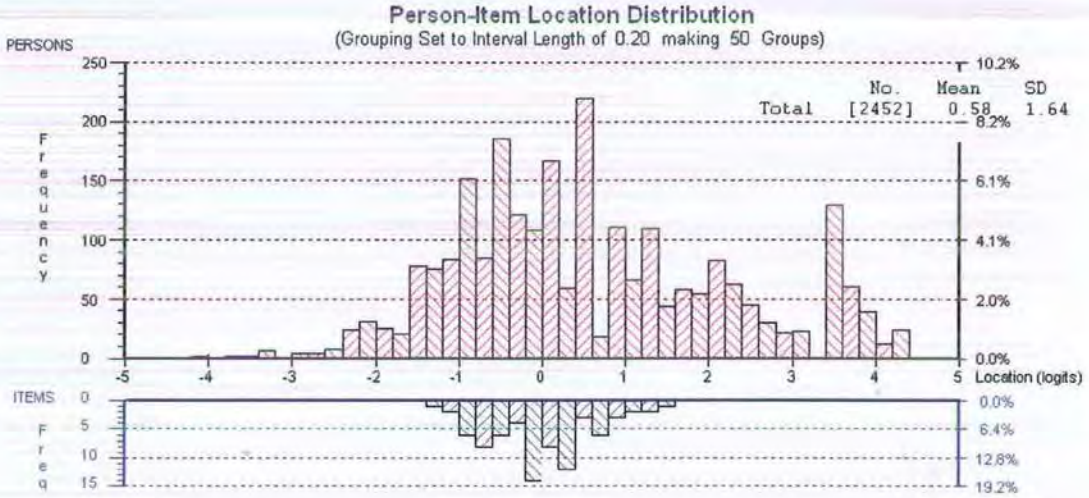


Figure 6.4 **Item locations and mathematics measures on the same scale**

Note on figure 6.4

1. The scale is in logits, the log odds of answering the response categories.
2. Mathematics measures from low to high are placed on the upper side of the scale and item locations (difficulties) from easy to hard are placed on the lower side of the scale.

Item Characteristic Curves

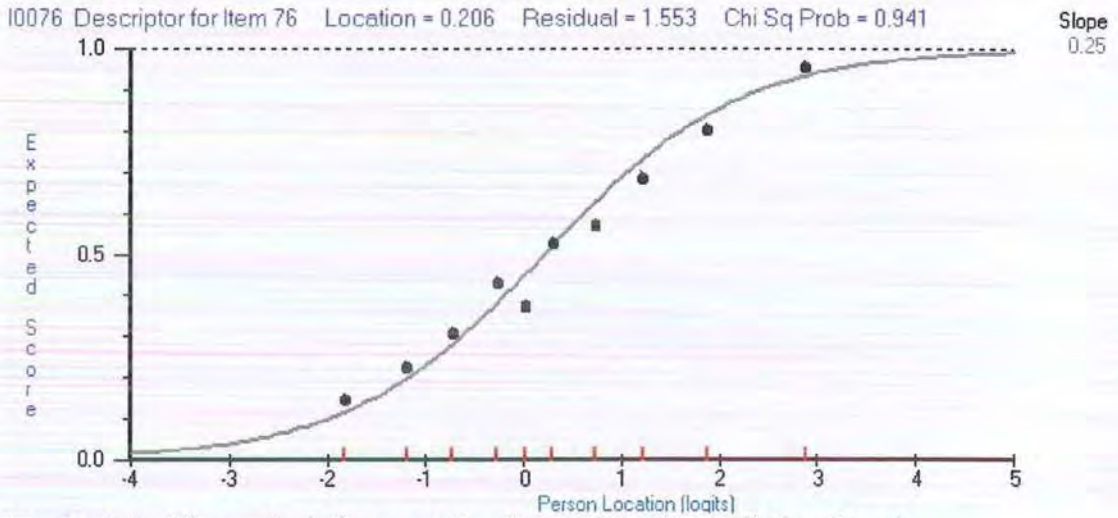


Figure 6.5 **Characteristic curve for item 76 (a Good-Fitting Item)**

The item characteristic curve for Item 76 (good-fitting item) of the mathematics scale is shown on Figure 6.6. The line indicates the expected score of mathematics ability groups, ranging from the lowest to highest ability groups, for each observed measure of a student ability group. When the observed scores closely follow the curve of expected values, the group is performing as expected on the item. Item 76 shows a good-fitting item to the model with all groups of mathematics ability close to the expected scores.

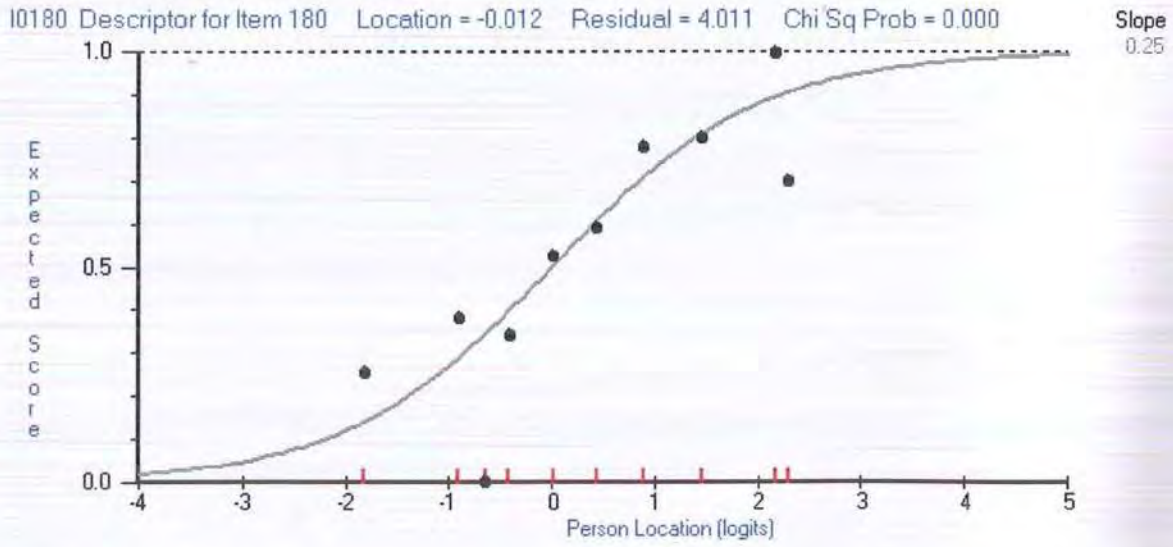


Figure 6.6 **Characteristic curve for item 180 (a Poor-Fitting Item)**

Item 180 is a not-so-good-fitting item of the mathematics scale. The item characteristic curve is shown on Figure 6.6. Four groups had higher expected scores and

two groups had lower than expected scores. This explains the poor fit of this item to the measurement model since many students did not perform as expected on this item.

Item Difficulties

After the Rasch analysis, the items were ordered in terms of their calibrated item difficulties (see Table 6.2 – 6.10) by sub-groups.

Table 6.2
Item difficulties for identification of equation from given choices (I=7, N=2,452)

Item Number	Item content	Difficulty
1(51)	Identification of an equation from given choices	-0.85
2(91)	Identification of an equation from given choices	-0.61
3(92)	Identification of both equations from given choices	-0.33
4(132)	Identification of both equations from given choices	-0.10
5(171)	Identification of an equation from given choices	+0.12
6(212)	Identification of both equations from given choices	+0.27
7(172)	Identification of both equations from given choices	+0.39

Notes on table 6.2

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.12 and 0.13 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items relating to the identification of the equation were found to be ordered from very easy (item 51) to moderately hard (item 172) (see Table 6.2). For example, the students found it very easy to identify the equations, item 51, item 91, and item 92. They found it easy (but harder) to identify the equations, item 132. For items 171, 212, and 172, they found it moderately hard to identify the equations, as would be expected.

Table 6.3
Item difficulties for identification of the true equation (I=11, N=2,452)

Item Number	Item content	Difficulty
1(214)	Identification of the true equation from given choices	-1.07
2(174)	Identification of the true equation from given choices	-0.83
3(53)	Identification of the true equation from given choices	-0.62
4(13)	Identification of the true equation from given choices	-0.05
5(74)	Selecting the true equation from given equations	+0.22
6(71)	Selecting the true equation from given equations	+0.68
7(134)	Identification of the true equation from given choices	+0.85
8(73)	Selecting the true equation from given equations	+0.85
9(112)	Selecting the true equation from given equations	+1.08
10(233)	Selecting the true equation from given equations	+1.37
11(34)	Selecting the true equation from given equations	+1.57

Notes on table 6.3

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items relating to the identification of the true equation were found to be ordered from very easy (item 214) to very hard (item 34) (see Table 6.3). For example, the students found that it very easy to identify the true equation from the given choices of item 214. They found it very easy (but harder) to identify the true equation of items 174 and 53, moderately easy for item 13, hard for item 74, and very hard for items 71, 134, 73, 112, 233 and 34, as would be expected.

Table 6.4
Item difficulties for identification of an equation with an unknown (I=3, N=2,452)

Item Number	Item content	Difficulty
1(16)	Identification of two equations with unknowns	-0.96
2(135)	Identification of an equation with an unknown	-0.71
3(14)	Identification of an equation with an unknown	-0.66

Notes on table 6.4

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the error is 0.13 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items on identifying equations with an unknown were found to be all very easy (items 16, 135 and 14) (see Table 6.4). The students found it very easy to identify the two equations with unknowns (item 16) and harder (but still very easy) to identify the equations with unknowns from items 135 and 14.

Table 6.5
Item difficulties for finding the true equation in different circumstances (I=8, N=2,452)

Item Number	Item content	Difficulty
1(59)	Finding the true equation when an unknown is replaced by 5	-1.27
2(138)	Finding the value of X which satisfies the equation $X \times 6 = 6$	-1.12
5(139)	Finding the true equation when an unknown replaced by 100	-0.93
4(220)	Finding the true equation when an unknown replaced by 79	-0.50
5(218)	Finding the value of an unknown which satisfies the equation $121 \div Y = 11$	-0.48
6(180)	Finding the true equation when an unknown replaced by 12	-0.01
7(83)	Finding the value of an unknown which satisfies the equation $Z \div 6 = 42$,	+0.74
8(97)	Finding the value of an unknown which satisfies the equation $\frac{Y}{5} = 60$	+1.37

Notes on table 6.5

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items on finding the value of an unknown that satisfies an equation were found to be ordered from very easy (item 59) to very hard (item 57) (see table 6.5). For example, the students found it very easy to find the true equation when an unknown is replaced by the number 5 (item 59). They found it harder (but still very easy) to find the value of X which satisfies the equation $X \times 6 = 6$ (item 138), the true equation when an unknown replaced by the number 100 (item 139), the true equation when an unknown is replaced by the number 79 (item 220) and the value of Y which satisfies the equation $121 \div Y = 11$ (item 218). They found it moderately easy (but harder) to find the true

equation when an unknown is replace by the number 12 (item 180). They found that it very hard to find the value of Z which satisfies the equation $Z \div 6 = 42$ (item 83) and the value of Y which satisfies the equation $\frac{y}{5} = 60$ (item 97).

Table 6.6
Item difficulties for finding the method to solve the equations (I= 17, N=2,452)

Item Number	Item Content	Difficulty
1(150)	Finding the method to solve the equation $J \div 65 = 130$	-0.28
2(109)	Finding the method to solve the equation $X \div 29 = 174$	-0.15
3(141)	Finding the method to solve the equation $P + 100 = 200$	-0.13
4(103)	Finding the method to solve the equation $96 + L = 386$	-0.12
5(68)	Finding the method to solve the equation $16 \times Q = 64$	-0.04
6(100)	Finding the method to solve the equation $X + 45 = 90$	-0.02
7(29)	Finding the method to solve the equation $Z \div 73 = 365$	+0.14
8(224)	Finding the method to solve the equation $56 + B = 168$	+0.19
9(106)	Finding the method to solve the equation $Z \times 35 = 140$	+0.24
10(104)	Finding the method to solve the equation $J - 35 = 105$	+0.24
11(185)	Finding the method to solve the equation $L - 47 = 188$	+0.27
12(223)	Finding the method to solve the equation $80 + F = 240$	+0.29
13(142)	Finding the method to solve the equation $75 + D = 375$	+0.37
14(140)	Finding the method to solve the equation $Y + 40 = 80$	+0.47
15(183)	Finding the method to solve the equation $125 + E = 250$	+0.54
16(182)	Finding the method to solve the equation $X + 61 = 122$	+0.73
17(60)	Finding the method to solve the equation $X + 100 = 100$	+0.95

Notes on table 6.6

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items relating to the identification of the method to solve the equations were found to be ordered from easy (item 150) to very hard (item 60) (see Table 6.6). Some examples are given now. Item 109 (Find the method to solve the equation X ÷ 29 = 174) and item 103 (Find the method to solve the equation 96 + L = 386) were found to be easy. Item 224 (Find the method to solve the equation 56 + B = 168) and item 104 (Find the method to solve the equation J – 35 = 105) were found to be of moderate difficulty. Item 183 (Find the method to solve the equation 125 + E = 250) and item 182 (Find the method to solve the equation X + 61 = 122) were found to be very difficult.

Table 6.7
Item difficulties for finding the solution of an equation (I=9, N=2,452)

Item Number	Item content	Difficulty
1(119)	Find the solution of Q × 24 = 168	-0.82
2(116)	Find the solution of Y + 14 = 140	-0.77
3(200)	Find the solution of 21 + Z = 63	-0.70
4(122)	Find the solution of 25 × F = 25	-0.62
5(241)	Find the solution of 7 + R = 84	-0.54
6(202)	Find the solution of 11 × D = 88	-0.45
7(157)	Find the solution of A – 10 = 100	-0.41
8(197)	Find the solution of M – 38 = 152	-0.06
9(117)	Find the solution of 175 = E – 5	+0.25

Notes on table 6.7

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items relating to finding the solutions to equations are ordered in difficulty from very easy (item 119) to moderately hard (item 117) (see Table 6.7). For example, the students found it very easy to find the solutions to the equations Q × 24 = 168 (item 119), Y + 14 = 140 (item 116), and 21 + Z = 63 (item 200), 25 × F = 25 (item 122), 7 + R = 84 (item 241), 11 × D = 88 (item 202), and A – 10 = 100 (item 157). They found it moderately easy to find the solution to the equation M – 38 = 152 (item 197) and they found it moderately hard to find the solution to the equation 175 = E – 5 (item 117).

The items relating to finding a solution to an equation involving a given condition (see Table 6.8) were found to be ordered from moderately hard (item 115) to very hard (item 164). For example, the students found it moderately hard to find the equation in which E has the highest value (item 115). They found it hard to find the value of X + 10, given X + 69 = 138 (item 76) , Y – 5, given Y × 7 = 49 (item 81), and to find the equation in which F is less than 90 by 6 (item 236). They found it very hard to find an equation which has the same solution as the equation C – 11= 22 (item 75), the value of E+10, given E × 12 = 60 (item 121), the value of X + 10, given X + 21 = 105 (item 156), and the value of B – 5, given B ÷ 5 = 60 (item 164).

Table 6.8
Item difficulties in order for finding the solution or equation which related to the given conditions (I=8, N=2,452)

Item Number	Item content	Difficulty
1(115)	Find the equation where E has the highest value	+0.08
2(76)	Find the value of X + 10, given X + 69 = 138	+0.21
3(81)	Find the value of Y – 5, given Y × 7 = 49	+0.28
4(236)	Find the equation which F is less than 90 by 6	+0.51
5(75)	Find the equation which is the same solution as the equation C – 11= 22	+0.63
6(121)	Find the value of E+10, Given E × 12 = 60	+0.65
7(156)	Find the value of X + 10, Given X + 21 = 105	+0.67
8(164)	Find the value of B – 5, Given B ÷ 5 = 60	+1.01

Notes on table 6.8

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items on selecting an equation converted from a verbal problem, or a problem converted from an equation, were found to be ordered from very easy (item 88) to hard (item 48) (see Table 6.9). Some examples are given now. The students found it very easy to select an equation which is converted from a verbal problem " Y students in a classroom were divided into 8 equal groups with 5 students each" (item 88). They found it easy to select an equation in finding out the value of X from a problem "Dang had X Baht and had 10 Baht more from selling eggs. The total sum of his money was 30 Baht." (item 45) and found it easy (but harder) to choose an equations which shows how many pieces of paper did Pooh collect from a problem "Pooh had 3 pieces of paper. She

collected Z pieces more. The total pieces were 20" (item 125). They found it moderately hard to select an equation which shows the total sum of John from a problem "John had the sum Y Baht. He bought a flashlight for 120 Baht and two bags for 70 Baht. 55 Baht remains." (item 129) and an equation which shows how many pieces did Adam buy from a problem "Adam bought Z pieces of pork, costing 3 Baht per each. The sum used was 54 Baht." (item 170). They found it hard to select a verbal problem which is related the equation $X \div 5 = 7$ (item 48).

Table 6.9

Item difficulties for selection an equation which is converted from a verbal problem or a problem which is converted from an equation (I=8, N=2,452)

Item Number	Item content	Difficulty
1(88)	Select an equation of the statement " Y students in a classroom was divided in to 8 equal groups with 5 students each.	-0.86
2(45)	Select an equation in finding out the value of X from a problem "Dang had X Baht and had 10 Baht more from selling eggs. The total sum of his money was 30 Baht."	-0.45
3(85)	Select an equation which shows how many items did Peter had solve more from a problem "Peter solved 5 items and solved Y more. In total, he solved 12 items".	-0.32
4(125)	Select an equation which shows how many pieces of paper did Pooh collect from a problem "Pooh had 3 pieces of paper. She collected Z pieces more. The total pieces were 20".	-0.21
5(126)	Select an equation which shows the temperature of yesterday from a problem "Today's temperature is 19c. Yesterday was Xc. The total temperatures were 41c".	-0.07
6(129)	Select an equation which shows the total sum of John from a problem "John had the sum Y Baht. He bought a flashlight for 120 Baht and two bags for 70 Baht. 55 Baht remains."	+0.10
7(170)	Select an equation which shows how many pieces did Adam buy from a problem "Adam bought Z pieces of pork, costing 3 Baht per each. The sum used was 54 Baht."	+0.14
8(48)	Select a verbal problem which is related the equation $X \div 5 = 7$	+0.22

Notes on table 6.9

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items on problem solving were found to be ordered from very easy (item 87) to moderately hard (item 90) (see Table 6.10). For examples, the students found it very easy to find the original amount from the problem "Dang had X Baht in his account and deposited 115 Baht more. The total was 321 Baht."(item 87). They found it

moderately easy (but harder) to solve the problems "The man has the cash of 4,650 Baht. After having its deposited in a bank, he has 3,500 Baht remaining. How much money did he deposit in a bank?" (item 249), and the problem "A teacher wants to divide 100 boy scouts in to equal groups with 9 members a group. Which keep one scout from group. How many groups will he divide?' (item 248).

Table 6.10
Item difficulties for problem solving (I=7, N=2,452)

Item Number	Item content	Difficulty
1(87)	Dang had X Baht in his account and deposited 115 Baht more. The total was 321 Baht. What is the original amount?	-0.66
2(249)	A man has cash of 4,650 Baht. After depositing some of it in a bank, he has 3,500 Baht remaining. How much money did he deposit in the bank?	-0.09
3(127)	One fence post is 180 cm. long. 40 cm. of the post is buried in the soil and Y cm. is above the soil. How many centimetres are above the soil?	-0.07
4(130)	Sopon wants to buy a 360 Baht slack. But he had only 180 Baht. How many flowers garlands does he have to sell to earn enough money if each garland costs 10 Baht?	-0.05
5(248)	A teacher wants to divide 100 boy scouts in to equal groups with 9 members a group. Which keep one scout from group. How many groups will he divide?	-0.03
6(47)	"A" had X Baht, "B" had 5 Baht more than "A".The total sum of the two was 65 Baht. How much money did "A" have?	+0.14
7(90)	A rope is M metres long. It is cut into 18 ropes with the length of 2 metres each. What is the length of the rope?	+0.16

Notes on table 6.10

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.11 and 0.14 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

They found it moderately hard to solve the problems "A" had X Baht, "B" had 5 Baht more than "A". The total sum of the two was 65 Baht. How much money did "A" have? (item 47) and the problem "A rope is M metres long. It is cut into 18 pieces with a length of 2 metres each. What is the length of the rope M?" (item 90).

Rasch Analysis Linked to the 78 Item Scale : 20 Items Scale

Further analysis with the RUMM program tested the extra 50 items (N=610) in order to create a linear scale of mathematics achievement with the 20 items that fitted the measurement model. Ten common items from the set of 78 items were included as part of the 50 items. The residuals were examined; the residuals being the difference between the expected item score calculated according to the Rasch measurement model and the actual item score of the students. This is converted to a standardized residual score in the computer program. The global item fit residuals and global student standardised fit residuals have a mean near zero and a standard deviation near one (see Table 6.11), indicating a reasonable fit to the measurement model. The probability of fit of items to the measurement model was then checked to identify items that fitted the model. The non-performing items of the mathematics achievement test (30 items out of 50) were deleted, thus creating a linear scale with only items that fitted the measurement model. Of the 20 items, 19 fitted the measurement model with probability $p>0.04$ (see Appendix D).

The item-trait test of fit examines the consistency of the item difficulties across the student mathematics measures along the scale. This determines whether there was agreement among student as to the difficulties of all items along the scale. The item-trait interaction was not statistically significant [Chi-square (df =160) =178.34, p =0.15]. This means that a unidimensional trait was measured.

The results of the analysis are set out in Tables 6.11 to 6.14, and Figures 6.7 to 6.12. Table 6.11 presents a summary of the global fit statistics of the measure of mathematics achievement, including the item-trait test of fit to the measurement model. The item difficulties in order for the 20 items are shown in Tables 6.12 to 6.14. Figure 6.7 shows the person measures of achievement and the item difficulties map for the mathematics test (20 items, 610 students), with the mathematics measures on the left hand side and the item difficulties on the right hand side. Figures 6.8 and 6.9 show response category curves for item 40 (good-fitting item) and item 43 (not-so-good fitting item). Figure 6.10 shows item locations on the lower side (LS) and mathematics measures on the upper side (US) on the same scale in logits. Figures 6.11 and 6.12 show the item characteristic curves for a good fitting item and a not-so-good-fitting item.

Appendix D shows, in probability order, the location on the continuum, fit to the measurement model and probability of fit to the model for the 20 items.

Table 6.11
Summary of fit statistics for mathematics achievement scale (I=20, N=610)

	Items	students
Number	20	610
Location mean	0.00	-0.57
Standard deviation	0.49	1.10
Fit statistic mean	0.71	0.15
Fit statistic standard deviation	0.89	0.60
Item-trait interaction chi square = 178.34		
Degrees of freedom = 160		
Probability of item-trait (p) = 0.15		
Student Separation Index = 0.76		
Power of test-of fit: Good (based on the Separation Index)		

Notes on Table 6.11

1. The item means are constrained to zero by the measurement model.
2. When the data fit the model, the fit statistics approximate a distribution with a mean near zero and a standard deviation near one The item fit and student fit are satisfactory, but not an excellent fit.
3. The item-trait interaction indicates the agreement displayed with all the items across all students from different locations on the scale (good for these data).This means that a unidimensional measure has been made.
4. The Student Separation Index is the proportion of observed student mathematics variance considered true (in this scale, 76% and is good.
5. Numbers are given to two decimal places because the errors are between 0.09 and 0.10 logits.

Person Separation Index

The Index of Person Separation (akin to traditional reliability) for the data of 20 items mathematics scale is 0.76 (see Table 6.11). This means that the proportion of observed variance considered true is 76 % and indicated that the student measures are well separated along the scale in comparison with the errors.

```
LOCATION          PERSONS                 ITEM DIFFICULTIES
            high measures     | hard items
                              |
                              |
                            X |
  4.0                         |
                              |
                              |
                              |
                              |
  3.0                      XX |
                              |
                              |
                              |
                            X |
  2.0                         |
                           XX |
                              |
                           XX |
                              |
  1.0                       X |
                           XX |
                          XXX |
                          XXX | 46  31  22
                        XXXXX | 32  43  50  23
  0.0                   XXXXXX | 37  24  45  41  36  40
                        XXXXXX | 42  20
                             X | 12  47
                  XXXXXXXXXXX | 18  17
              XXXXXXXXXXXXXXXX |
 -1.0   XXXXXXXXXXXXXXXXXXXX |
              XXXXXXXXXXXXXXX |
                         XXX |
                   XXXXXXXXXX |
                             X | 7
 -2.0                XXXXXXXX |
                              |
                         XXX |
                              |
                              |
 -3.0                         |
                            X |
                              |
                              |
                              |
 -4.0     low measures        |  easy items
-------------------------------------------------------
```

Figure 6.7 **Person measures of achievement and item difficulty map for the mathematics test (N=610, I=20).**

Notes on Figure 6.7

1. The scale is in logits, the log odds of answering positively.
2. Mathematics measures are calibrated on the same scale as the item difficulties.
3. Measures are ordered from low to high on the LHS and item difficulties are ordered from easy to hard on the RHS.
4. Items at the easy end of the scale are answered positively by most students. As the items become harder, students need a higher mathematics measure to answer the items positively.
5. Each x represents 5 students.

Figure 6.7 shows that targeting of the item difficulties is not as good as it should be. There are no items matching persons at either the low end of the scale (-0.8 to -3.2 logits) and to the high end of the scale (+0.5 to +4.2 logits), indicating the improvements in the Person Separation Index may be possible for the test if both easy items or hard items are added.

Category Response Curves

The RUMM program provides a category response curve for each item to check whether the category responses are being answered consistently and logically. A perusal of the category response curves for the 20 items indicates that the students answered the response categories consistently and logically. The items contained two response categories earning 0 and 1 mark. For example, Figure 6.8 shows the category response curve for the good fitting item (item 40). The category 0 means 0 mark (category response wrong) and category 1 means 1 mark (category response right).

Item 40 is a good-fitting item with a chi square probability of 0.99. Its difficulty is 0.19 and this means that the students found that the item is relatively hard for them. Figure 6.8 shows that the category curve 0 (category response wrong) indicates that when a student has very low mathematics measure (-3.0 logits), then the probability of answering in this category (getting 0) is 0.98 (very high as expected). As the student's mathematics measure increases (to –1.0 logits), then the probability of getting 0 drops to near 0.78 (as expected). If the student's mathematics measure increases (to 0.0 logits), then the probability of getting 0 drops to near 0.57 (as expected).When the student's mathematics measure increases to+3.0 logits, then the probability of getting 0 drops to zero (as expected).

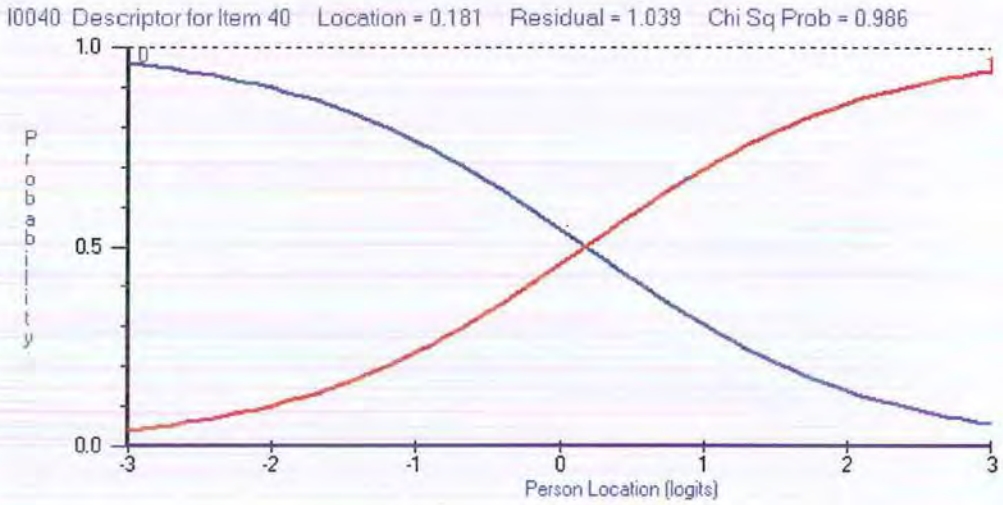


Figure 6.8 **Response category curve for item 40 (good-fitting item)**

For curve 1 (category response right), when the student has a very low mathematics measure (-3 logits), then the probability of answering right (getting 1) is near zero (very low as expected). When the student mathematics measure increases to -1 logits, then the probability of answering right (getting 1) increases to 0.23 (as expected). When the student mathematics measure increases to 0 logits, then the probability of answering right increases to near 0.50 (as expected). When the student mathematics measure increases to +3, the probability of answering right increases to 1 (very high as expected).

Item 43 is a hard item that doesn't fit the measurement model as well as one would like. Nevertheless, the response category curve is good. It has a difficulty of 0.31 on this scale, which indicates that many students found that it was rather hard for them. Figure 6.9 shows that the category curve 0 (category response wrong) indicates that when a student has a very low mathematics measure (-3.0 logits), then the probability of answering in this category (getting 0) is 0.98 (very high as expected). As the student's mathematics measure increases (to –1.0 logits), then the probability of getting 0 drops to near 0.82 (as expected). If the student's mathematics measure increases (to 0.0 logits), then the probability of getting 0 drops to near 0.62 (as expected). When the student's mathematics measure increases to+3.0 logits, then the probability of getting 0 drops to zero (as expected).

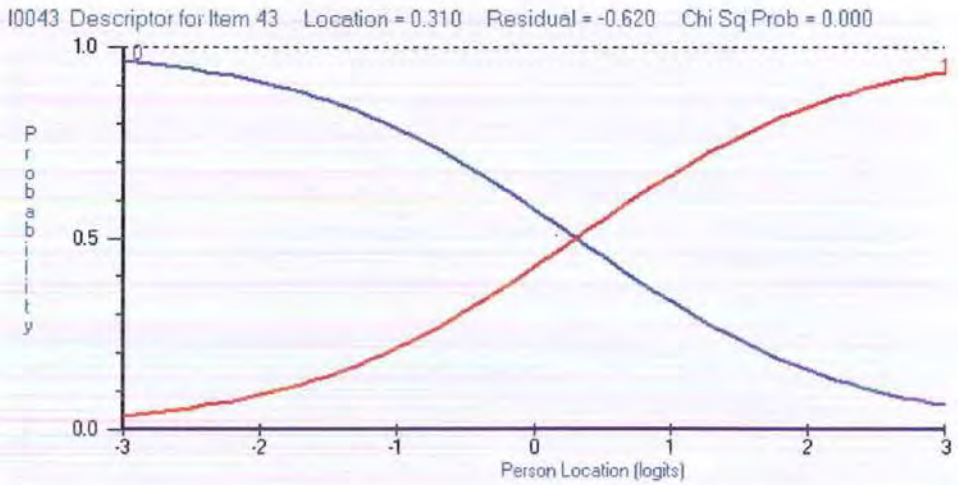


Figure 6.9 **Response category curve for item 43 (not-so-good-fitting item)**

For curve 1 (category response right), when the student has a very low mathematics measure (-3.0 logits), then the probability of answering right (getting 1) is near zero (very low as expected). When the student mathematics measure increases to –1.0 logits, then the probability of answering right (getting 1) increases to 0.22 (as expected). When the student mathematics measure increases to 0.0 logits, then the probability of answering right increases near 0.5 (as expected). When the student mathematics measure increases to +4.0 logits, the probability of answering right increases to 1 (as expected).

Targeting

The locations (difficulties) of the items cover the middle range of mathematics measures, but not the lower and higher ranges, as well as they could (see Figure 6.10). The difficulties of the items range from about -1.7 to +0.5 logits and cover part of the range of mathematics measures (about –4.1 to +4.4 logits, see Figure 6.10). This means that the targeting of the mathematics items could be improved, and easier and harder items could be added in a revision of the scale.

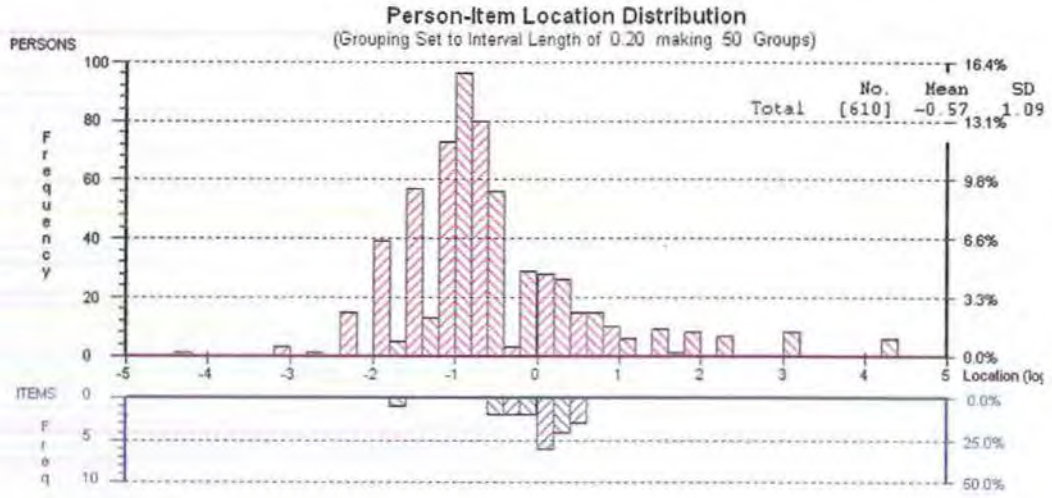


Figure 6.10 **Item locations and mathematics measures on the same scale**

Note on figure 6.10

1. The scale is in logits, the log odds of answering the response categories.
2. Mathematics measures from low to high are placed on the upper side of the scale and item locations (difficulties) from easy to hard are placed on the lower side of the scale.

Item Characteristic Curves

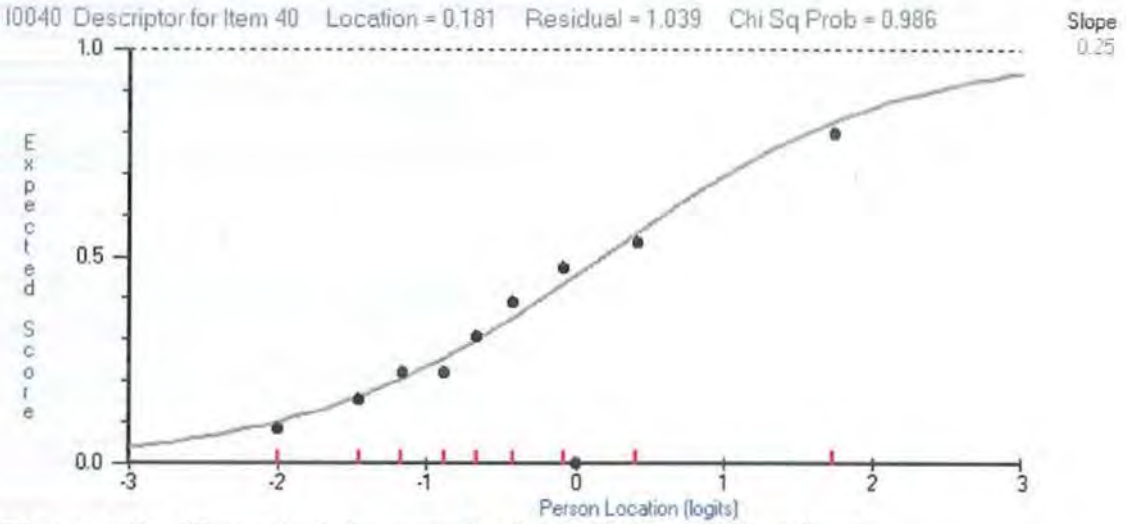


Figure 6.11 **Characteristic curve for item 40 (good fitting item)**

The item characteristic curve for Item 40 (good fitting item) of the mathematics scale is shown on Figure 6.11. The line indicates the expected score of mathematics ability groups, ranging from the lowest to highest ability groups. Each black dot represents the observed score of a student ability group. When the observed scores closely follow the curve of expected values, the group is performing as expected on the

item. Item 40 shows good fitting item to the model with all groups of the mathematics ability close to the expected scores.

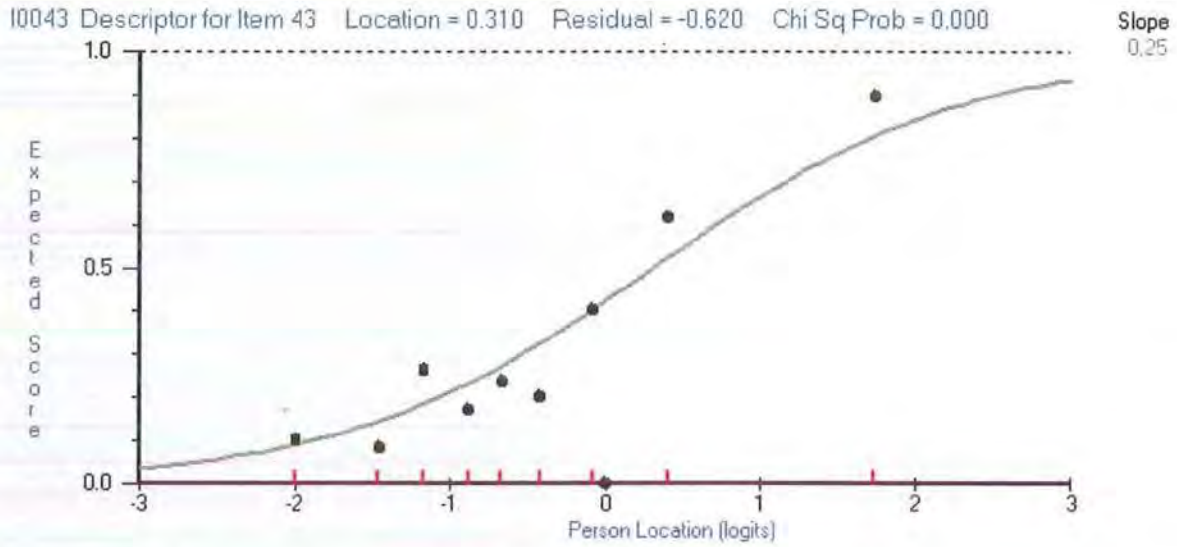


Figure 6.12 **Characteristic curve for item 43 (poor fitting item)**

Item 43 is a not-so-good-fitting item of the mathematics scale and its item characteristic curve is shown on Figure 6.12. The line indicates the expected score of mathematics ability groups, ranging from the lowest to highest ability groups. Each black dot represents the observed score of a student ability group. When the observed scores closely follow the curve of expected values, the group is performing as expected on the item. Item 43 has only 2 from 10 (20%) mathematics ability groups close to the expected scores.

Item Difficulties

After the Rasch analysis, the items were ordered in terms of their calibrated item difficulties (see Tables 6.12 – 6.14).

Table 6.12
Item difficulties for finding the solution to an equation (I=20, N=610)

Item Number	Item content	Difficulty
1(18)	Find the solution of $\frac{2X}{8} + \frac{5X}{20} = 6$	-0.53
2(17)	Find the solution of $Y - 9 = \frac{1}{4}$	-0.46
3(20)	Find the solution of $\frac{X+2}{6} = \frac{2X-1}{6}$	-0.11

Notes on table 6.12

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the error is 0.09 logots.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items on finding the solution to an equation were found to be all easy (items 18, 17 and 20) (see Table 6.12). The students found it easy to find the solutions to the equations $\frac{2X}{8} + \frac{5X}{20} = 6$ (item 18) and the equation $Y - 9 = \frac{1}{4}$ (item 17). They found it easy (but harder) to find the solution to the equation $\frac{X+2}{6} = \frac{2X-1}{6}$ (item 20).

Table 6.13
Item difficulties for finding the solution or equation which related to the given condition (N=610)

Item Number	Item content	Difficulty
1(47)	Find the equation which has the same solution as the equation $5X + 10 = 40$	-0.23
2(42)	Find the equation where Y is equal to 10	-0.15
3(37)	Find the different solution between the equations $3Y-6=Y+4$ and $2Y - 5 = 35$	+0.09
4(24)	Find the value of 2X – 14, Given X – 7 = 3	+0.11
5(45)	Find the equation which X equal to 18	+0.12
6(41)	Find the value of X which makes $\frac{X}{6} + \frac{5X}{6} = 3$ less than $2X - 5 = 43$	+0.12
7(36)	Find the value of X which makes $5X = 4X + 9$ more than $5X + 3 = 3X + 9$	+0.15
8(40)	Find the value of X which makes $5X - 5 = 30$ more than $5X = 3X + 6$?	+0.18
9(32)	Find the value of $4X - 2$ Given $6X - 24 = 6 - 4X$	+0.28
10(50)	Find the choice which has two equations in which choice have the same solutions	+0.31
11(43)	Find the equation which has different solution from others	+0.31
12(23)	Find the equation which has the least solution	+0.39
13(46)	Find the equation which has the solution more than 40	+0.42
14(31)	Find the value of 3A+3B, Given A=4-B	+0.43
15(22)	Find the equation which has the least value of X	+0.51

Notes on table 6.13

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.09 and 0.10 logits.
3. Items are ordered from Find the value of 2X – 14, Given X – 7 = 3 easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The items relating to finding a solution to an equation which relating to a given condition (see Table 6.13) were found to be ordered from moderately easy (item 47) to moderately hard (item 22). For examples, the students found it moderately easy to find the equation which has the same solution as the equation $5X+10=40$ (item 47) and the equation which Y equal to 10 (item 42). They found it moderately hard to find the different solution between the equations $2Y-5=35$ and $2Y - 5 = 35$ (item 37), the

value of the equation 2X – 14, Given X – 7 = 3 (item 24), the equation which X equal to 18 (item 45), the value of X which makes $\frac{X}{6} + \frac{5X}{6} = 3$ less than $2X - 5 = 43$ (item 41), the value of X which makes $5X = 4X + 9$ more than $5X + 3 = 3X + 9$ (item 36), and the value of X which makes $5X - 5 = 30$ more than $5X = 3X + 6$ (item 40). They found it hard to find the value of x, Given $6X - 24 = 6 - 4X$ (item 32) and the two equations in the choice have the same solutions (item 50). They also found it hard to find the equation which has different solution from the others equations (item 43), and the equation which has the least solution (item 23). They found it moderately hard to find the equation which has the solution more than 40 (item 46), the value of 3A + 3B, Given A = 4 - B (item31), and the equation which has the least value of X (item22).

Table 6.14
Item difficulties for selection an equation which is converted from a verbal problem (I=2, N=610)

Item Number	Item content	Difficulty
1(7)	Select the equation which shows the total sum of John from a problem "John had the sum Y baht. He bought a flashlight for 120 baht and two bags for 70 baht. 55 baht remains."	-1.65
8(12)	Find the equation of the statement " A had X baht. B had 50 baht more two times of A. The sum of the two equals to four times of A's"	-0.29

Notes on table 6.14

1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the error is .09 logits.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The students found it extremely easy to select an equation which is converted from the problem "John had the sum Y Baht. He bought a flashlight for 120 Baht and two bags for 70 Baht and 55 Baht remains. What is the total sum that John had?" (item 7). They found it moderately easy to select an equation which is converted from the problem "A had X Baht. B had 50 Baht more two times of A. The sum of the two equals four times of As"(item 45).

Summary

The computer program RUMM (Andrich et al., 2003) was very useful in analysing data on mathematics achievement tests. The Rasch analysis showed that:

1. 78 of the original 250 items of test 1 to test 6 and 20 of the 50 items of test 7 fitted the measurement model with a probability >0.04;

2. There was good global item fit to the measurement model;

3. Global person fit to the measurement model was acceptable;

4. The item-trait interaction chi-squares were not statistically significant, indicating that a uni-dimensional trait was measured (or at least a dominant trait was present);

5. The Student Separation Indices were 0.83 for the first analysis and 0.76 for the second analysis, indicating that the errors were small in relation to the separation of persons along the scales;

6. The targeting of the item locations against the student measures needed improvement and, in any revision of the scale for these students, some harder items have to be added to cater for those with higher mathematics abilities and some easy items to cater for those of lower mathematics abilities.

The evidence shows that a reliable scale was constructed with the 98 items (78+20) from which valid inferences could be drawn.

The item bank of mathematics on equations for the year 6 (Prathom Suksa 6) students contained 98 items which fitted the measurement model and consisted of:

1. Seven items relating to the identification of an equation, ordered from very easy (difficulty = -0.85) to moderately hard (difficulty = +0.39);

2. Eleven items relating to the identification of the true equation, ordered from very easy (difficulty = -1.07) to very hard (difficulty = +1.57);

3. Three items on identifying equations with an unknown, were all very easy (difficulties from -0.96 to -0.66);

4. Eight items on finding the value of an unknown that satisfies the equation, ordered from very easy (difficulty = -1.27) to very hard (difficulty = +1.37);

5. Seventeen items relating to Identify the Method to solve the Equation, ordered from very easy (difficulty = -0.28) to extremely hard (difficulty = +0.95);

6. Twelve items relating to finding the solutions to equations, ordered from very easy (difficulty = -0.82) to moderately hard (difficulty = +0.25);

7. Twenty-three items relating to finding a solution of an equation which related the given condition, ordered from moderately easy (difficulty = -0.23) to very hard (difficulty = +1.01);

8. Nine items on selecting an equation converted from a verbal problem or a verbal problem related to an equation, ordered from very easy (difficulty = -0.86) to hard (difficulty = +0.22);

9. Seven items on problem solving, ordered from very easy (difficulty = -0.66) to moderately hard (difficulty = +0.16).

The next chapter (Chapter Seven) describes the data analysis for the computerized adaptive test designed by the researcher for Year 6 (Prathom Suksa 6) Primary School students taking mathematics on equations in Thailand.

CHAPTER 7

DATA ANALYSIS (PART II)
THE COMPUTERIZED ADAPTIVE TESTING RESULTS

This chapter contains a description of the results for the computerized adaptive testing, using a computer program designed by the author. The SPSS computer program (Pallant, 2001) was used to analyse data from 400 Prathom Suksa 6 students. The frequencies and percentages of mathematics ability were used as the indicators to examine mathematics competencies of the students. A one-way ANOVA was used to examine differences in test length and testing times among the different groups relating to stopping criteria and mathematics competencies, and also to examine differences in mathematics competencies among the different groups of stopping criteria. ANOVA is the appropriate statistic to use because there are more than two groups of the students and test length, testing times and because mathematics competencies were measured on ratio or interval scales (Cavana et al., 2001). Because the F statistics were significantly different, the Sheffe Multiple Range test was used to determine between which groups the true differences lie (Cavana et al., 2001). The frequency of Mathematics competencies of the students, one-way ANOVA, and Sheffe Multiple Range test results are shown through tables and descriptive text. The presentation begins with a description of the analysis for the mathematics achievement that is reported for 400 Prathom Suksa 6 students. The ANOVA and the Sheffe Multiple Range are used to show the mean differences in test length and testing times, among stopping criteria and mathematics competencies, and the mean difference of mathematics competencies for different groups of stopping criteria. A summary list of the main findings is presented at the end of the chapter.

Mathematics Competency

The result of the analysis of mathematics competencies of Prathom Suksa 6 students is set out in Table 7.1. It presents frequencies and percentages of mathematics competencies of the students in the three groups (low, moderately high, and high).

Table 7.1
Frequency table for mathematics competencies

Achievement	Frequency	Percent	Cumulative Percent
Low	67	16.75	16.75
Moderately high	301	75.25	92.00
High	32	8.00	100.00
Total	400	100.00	

As can be seen from Table 7.1, the results showed that there are 67 (16.75%) students and they can be regarded as having a low mathematics achievement (mathematics measures were from -1.02 to 0.00 logits). From 0.00 to +1.00 logits, there are 301 (75.25%) students and they can be regarded as having a moderately high mathematics achievement. From +1.00 to +3.00 logits, there are 32 (8.00%) students and they can be regarded as having a high mathematics achievement.

Differences in Test Length and Testing Times Among Different Groups by Stopping Criteria and Mathematics Competencies

The results of the analysis of the test relating to different test lengths and testing times among four groups for stopping criteria [see description of stopping criteria on page 9] and three groups of mathematics competencies of the students with the Mathematics Computerized Adaptive Testing are set out in Tables 7.2 to 7.9. Tables 7.2, 7.4, 7.6, and 7.8 show the F values to examine the difference in test length and testing times among the different groups for stopping criteria and mathematics competencies, while the Sheffe Multiple Range test results for the differences are set out in Tables 7.3, 7.5, 7.7, and 7.9.

Table 7.2
Test length for the different groups by stopping criteria

Source of Variation	Sum of Squares	df	Mean Square	F	p
Between Groups	1560.31	3.00	520.10	191.30	.00*
Within Groups	1076.63	396.00	2.72		
Total	2636.94	399.00			

Note

1. p means significance based on the F value. * $p < 0.05$.

As can be seen from Table 7.2, the F test shows that the difference in the means of the students for the four groups by stopping criteria, $SEE \leq 0.20$ (1), $SEE \leq 0.30$ (2), $SEE \leq 0.40$ (3), and $SEE_m - SEE_{m-1} \leq 0.005$ (4), were significantly different at the 5 per cent significance level in regards to test length (F = 191.30, df = 3, 396, p = 0.00). That is, there were significant differences in the mean test length levels of students in the four groups by stopping criteria.

To determine between which groups test lengths are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 7.3.

Table 7.3
Differences in test length by stopping criteria

Stopping criteria	Mean	(2) 3.14	(1) 4.34	(4) 6.83	(3) 8.14
$SEE \leq 0.30$ (2)	3.14		1.20*	3.69*	5.00*
$SEE \leq 0.20$ (1)	4.34			2.49*	3.80*
$SEE_m - SEE_{m-1} \leq 0.005$ (4)	6.83				1.31*
$SEE \leq 0.40$ (3)	8.14				

Note
* The mean difference was significant at the 0.05 level.

As can be seen from Table 7.3, the results showed that mean test length (number of items) for the four groups by stopping criteria was 4.34 for the first criteria, 3.14 for the second, 8.14 for the third, and 6.83 for the fourth. There were six main points of difference in test length by stopping criteria. The third group was significantly different from groups 2, 1, and 4 at p=0.05; the fourth group was significantly different from groups 2 and 1 at p=0.05; and the first group was significantly different from group 2 at p=0.05.

Table 7.4
Testing times for the different groups by stopping criteria

Source of Variation	Sum of Squares	df	Mean Square	F	p
Between Groups	755.89	3.00	251.96	53.85	.00*
Within Groups	1,852.78	396.00	4.68		
Total	2,608.67	399.00			

Note

1. p means significance based on the F value. * $p < 0.05$.

As can be seen from Table 7.4, the F test shows that the difference in the means of the students for the four groups by stopping criteria, $SEE \leq 0.20$ (1), $SEE \leq 0.30$ (2), $SEE \leq 0.40$ (3), and $SEE_m - SEE_{m-1} \leq 0.005$ (4), were significantly different at the 5 per cent significance level, in regards to testing time ($F = 53.85$, $df = 3, 396$, $p = 0.00$). That is, there were significant differences in the mean testing time levels of students in the four groups by stopping criteria.

To determine between which groups, testing times are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 7.5.

Table 7.5
Differences in testing time by stopping criteria

Stopping criteria	Mean	(2) 2.38	(1) 3.33	(4) 5.26	(3) 5.74
$SEE \leq 0.30$ (2)	2.38		0.95*	2.88*	3.36*
$SEE \leq 0.20$ (1)	3.33			1.93*	2.41*
$SEE_m - SEE_{m-1} \leq 0.005$ (4)	5.26				0.48
$SEE \leq 0.40$ (3)	5.74				

Note
* The mean difference was significant at the 0.05 level.

As can be seen from Table 7.5, the results showed that mean testing times for the four groups by stopping criteria was 3.33 minutes for the first criteria, 2.38 minutes for the second, 5.74 minutes for the third, and 5.26 minutes for the fourth. There were five main points of difference in testing times by stopping criteria. The third group and the fourth group were significantly different from groups 2 and 1 at p=0.05; and the first

group was significantly different from group 2 at p=0.05. The third group with the stopping criteria of $SEE \leq 0.40$ was not significantly different from group 4.

Table 7.6
Test length for the different groups of mathematics competency

Source of Variation	Sum of Squares	df	Mean Square	F	p
Between Groups	207.04	2	103.52	16.91	.00*
Within Groups	2429.89	397	6.12		
Total	2636.94	399			

Note

1. p means significance based on the F value. * $p < 0.05$.

As can be seen from Table 7.6, the F test shows that the difference in the means of the students in the different groups of mathematics competency, high, moderately high, and low, were significantly different at p=0.05 (at the 5 per cent significance level) in regards to test length (number of items) ($F = 16.91$, $df = 2, 397$, $p = 0.00$). That is, there were significant differences in the mean test length levels of students in the three groups of the mathematics competency.

To determine between which groups, test lengths are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 7.7.

Table 7.7
Differences in test length by mathematics competencies

Mathematics competencies		(2)	(3)	(1)
	Mean	5.21	6.31	7.07
Moderately high (2)	5.21		1.10	1.86*
High (3)	6.31			0.76
Low (1)	7.07			

Note

* The mean difference was significant at the 0.05 level.

As can be seen from Table 7.7, the results showed that mean test length (number of items) for the three groups of mathematics competency was 7.07 for the first (low), 5.21 for the second (moderately high), and 6.31 for the third (high). The test length of the students in the first group with the low mathematics competency was significantly different from that in group2 (moderately high) at the 5 per cent significance level.

Table 7.8
Testing times for the different groups by mathematics competency

Source of Variation	Sum of Squares	df	Mean Square	F	p
Between Groups	52.40	2	26.20	4.07	0.02*
Within Groups	2556.27	397	6.44		
Total	2608.67	399			

Note

1. p means significance based on the F value. * $p < 0.05$.

As can be seen from Table 7.8, the F test shows that the difference in the means of the students in the different groups of, high, moderately high, and low mathematics competency, were significantly different. at the 5 per cent significance level, in regards to testing times ($F = 4.07$, $df = 2, 397$, $p = 0.02$). That is, there were significant differences in the mean testing time levels of students in the three groups of the mathematics competency.

To determine between which groups, testing times are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 7.9.

Table 7.9
Differences in testing time by mathematics competencies

Mathematics competencies		(2)	(30	(1)
	Mean	3.97	4.50	4.92
Moderately high (2)	3.97		0.53	0.94*
High (3)	4.50			0.41
Low (1)	4.92			

Note

* The mean difference was significant at 0.05 level.

As can be seen from Table 7.9, the results showed that mean testing time for the three groups of mathematics competency was 4.92 minutes for the first (low), 3.97 minutes for the second (moderately high), and 4.50 minutes for the third (high). There was one main point of difference in testing time by mathematics competencies. The first group with the low mathematics competency was significantly different from group 2 (moderately high) at the 5 per cent significance level.

Differences in Mathematics Competencies Among Different Groups by Stopping Criteria

The results of the analysis of the test of different mathematics competencies among four groups for stopping criteria of Mathematics Computerized Adaptive Testing are set out in Tables. 7.10 and 7.11. Table 7.10 shows the F values to examine the difference in mathematics competencies among the different groups for stopping criteria, while the Sheffe Multiple Range test results for the differences are set out in Table 7.11.

Table 7.10
Mathematics competencies for the different groups by stopping criteria

Source of Variation	Sum of Squares	df	Mean Square	F	p
Between Groups	3.41	3.00	1.14	5.09	.00*
Within Groups	88.46	396.00	.22		
Total	91.87	399.00			

Note

1. p means significance based on the F value. * $p < 0.05$.

As can be seen from Table 7.10, the F test shows that the difference in the means of the students for the four groups by stopping criteria, $SEE \leq 0.20$ (1), $SEE \leq 0.30$ (2), $SEE \leq 0.40$ (3), and $SEE_m - SEE_{m-1} \leq 0.005$ (4), were significantly different at the 5 per cent significance level in regards to mathematics competency ($F = 5.09$, $df = 3, 396$, $p = 0.00$). That is, there were significant differences in the mean mathematics competency levels of students in the four groups by stopping criteria.

To determine between which groups mathematics competencies are significantly different, the Sheffe Multiple Range test was performed. The results are shown in Table 7.11.

Table 7.11
Differences in mathematics competency by stopping criteria

Stopping criteria	Mean	(2) 0.38	(3) 0.54	(1) 0.57	(4) 0.63
$SEE \leq 0.30$ (2)	0.38		0.16	0.19*	0.25*
$SEE \leq 0.40$ (3)	0.54			0.03	0.09
$SEE \leq 0.20$ (1)	0.57				0.06
$SEE_m - SEE_{m-1} \leq 0.005$ (4)	0.63				

Note
* The mean difference was significant at the 0.05 level.

As can be seen from Table 7.11, the results showed that mean mathematics competency for the four groups by stopping criteria was 0.57 logits for the first criteria, 0.38 logits for the second, 0.54 logits for the third, and 0.63 logits for the fourth. There were two main points of difference in mathematics competency by stopping criteria. The second group was significantly different from group 1 and group 4 at p=0.05.

Summary of Results

The main findings are summarised.

Mathematics Competencies

There were 72.25 %, 16.75%, and 8% of the Prathom Suksa 6 students having a moderately high, low, and high mathematics achievement respectively.

Test Length ,Testing Times and Mathematics Competencies in Different Groups by Stopping Criteria

The four groups of stopping criteria were $SEE \leq 0.20$ (group 1), $SEE \leq 0.30$ (group 2) $SEE \leq 0.40$ (group 3) and $SEE_m - SEE_{m-1} \leq 0.005$ (group 4).

1. Test lengths were significantly different at p=0.05 among four groups of stopping criteria (F = 191.30, df = 3, 396, p = 0.00).
2. The mean highest test length (8.14 items) and the mean lowest test length (3.14 items) were in group 3 (stopping criteria is $SEE \leq 0.40$) and group 2 (stopping criteria is $SEE \leq 0.30$). Each group was significantly different at p=0.05 from the others.
3. Testing times were significantly different at p=0.05 among the four groups of stopping criteria (F = 53.85, df = 3, 396, p = 0.00).
4. The mean highest testing time (5.74 minutes) and the mean lowest testing time (2.38 minute) were in group 3 (stopping criteria is $SEE \leq 0.40$) and group 2 (stopping criteria is $SEE \leq 0.30$). Each group was also significantly different at p=0.05 from the others.
5. Mathematics competencies were significantly different at p=0.05 among the four groups of stopping criteria (F = 5.09, df = 3, 396, p = 0.00).
6. The mean highest mathematics competency (0.63 logits) and the mean lowest mathematics competency (0.38 logits) were in group 4 (stopping criteria is $SEE_m - SEE_{m-1} \leq 0.005$) and group 2 (stopping criteria is $SEE \leq 0.30$). Students mathematics competency in group 2 (stopping criteria is $SEE \leq 0.30$) was significantly different from group 1(stopping criteria is $SEE \leq 0.20$) and group 4 (stopping criteria is $SEE_m - SEE_{m-1} \leq 0.005$) at p=0.05.

Test Length and Testing Times in Different Groups by Mathematics Competencies

The three groups of mathematics competencies were low (group1), moderately high (group 2), and high (group 3).

1. Test lengths were significantly different at p=0.05 among three groups of mathematics competencies (F = 16.91, df = 2, 397, p = 0.00).
2. The mean highest test length (7.07 items) and the mean lowest test length (5.21 items) were in group 1 (low mathematics competency) and group 2 (moderately high mathematics competency). There was an only one significantly different test length at p=0.05 between students in group 1 (low mathematics competency) and group 2 (moderately high mathematics competency).
3. Testing times were significantly different at p=0.05 among the three groups of mathematics competencies (F = 4.07, df = 2, 397, p = 0.02).
4. The mean highest testing time (4.92 minutes) and the mean lowest testing time (3.97 minutes) were in group 1 (low mathematics competency) and group 2 (moderately high mathematics competency). There was an only one significantly different testing times at p=0.05 between students in group 1 (low mathematics competency) and group 2 (moderately high mathematics competency).

The next chapter (Chapter Eight) describes the data analysis for measures of student attitudes towards mathematics computerized adaptive testing.

CHAPTER 8

DATA ANALYSIS (PART III)
RASCH MEASUREMENT OF STUDENT ATTITUDES

This chapter presents the Rasch analysis results for student attitudes to a Computerized Adaptive Test. The RUMM computer program (Andrich et al., 2003) was used to investigate fit to the measurement model and to calibrate item difficulties and student measures on the same scale. The presentation begins with a description of the analysis for the attitude that is reported for 30 items. The Rasch analysis provides data on global item and person fit to the measurement model, item thresholds, individual item fit, dimensionality, reliability, Student Separation Index, Item Characteristic Curves and Response Category Curves, and targeting. The meaning of the attitude scale is discussed and a summary list of the main findings is presented at the end of the chapter.

Rasch Analysis

Overall Comment

Initial analysis with the RUMM program tested the 30 items (N=400) in order to create a linear scale of student attitudes to a Computerized Adaptive Test. The item thresholds were checked so that only those items with ordered thresholds (indicating that the response categories for the item were answered consistently and logically) were included in the final analysis. That meant that students who answered the neutral category were deleted leaving four response categories (and corresponding three thresholds). After that, the residuals were examined; the residuals being the difference between the expected item score calculated according to the Rasch measurement model and the actual item score of the students. This is converted to a standardized residual score in the computer program. The global item fit residuals and global student fit residuals have a mean near zero and standard deviation near one (see Table 8.1), indicating a reasonable fit to the measurement model. The probability of fit of items to the measurement model was then checked to identify items that fitted the

model. Of the 30 items, 27 fitted the measurement model with probability p>0.04 and there was a very good overall fit (see Appendix F and Table 8.1).

The item-trait test of fit examines the consistency of the item difficulties across the student attitude measures along the scale. This determines whether there was agreement among student as to the difficulties of all items along the scale. The item-trait interaction was not statistically significant [Chi-square (df =150) =165.40, p =0.18]. This means that a unidimensional trait was measured.

In Rasch analysis, the items are designed in a conceptual order by difficulty and this order was tested as satisfactory. The data for the items have to also fit the measurement model in order to create a linear scale and this was tested as satisfactory. The person measures and item difficulties were calibrated on the same scale by the RUMM 2010 program, thus providing the creation of a linear measure of student attitude towards Computerized Adaptive Testing.

The results of the analysis are set out in Tables 8.1 to 8.7, and Figures 8.1 to 8.7. Table 8.1 presents a summary of the global fit statistics of the measure of students' attitude towards Computerized Adaptive Testing (CAT), including the item-trait test of fit to the measurement model. An example of item thresholds for the attitude scale is shown in Table 8.2. The thresholds are ordered from low to high in line with the ordering of the response categories which indicates that students answered the response categories consistently and logically. The number of response categories for the attitude scale was four, which included strongly disagree, disagree, agree, and strongly agree, and were scored 0, 1, 2, and 3 respectively. Tables 8.3 to 8.7 show the item difficulties in order for five sub-groups: (1) Like and Interest in CAT; (2) Confidence with and Use of CAT; (3) CAT as Modern and Useful; (4) CAT as Reliable, Fair and Good; and (5) CAT Recommendations. Figure 8.1 shows a graph of the scale of attitude towards Computerized Adaptive Testing of students (30 items, 3 thresholds) for the 400 students, with the attitude measures on the left hand side and the thresholds on the right hand side. Figure 8.2 to 8.3 show response category curves for item 19 (good-fitting item) and item 9 (not-so-good fitting item). Figure 8.4 shows the items thresholds on the lower side (LS) and attitude measures on the upper side (US) on the same scale in logits. Figure 8.5 shows item locations (LS) and attitude measures (US) on the same scale in logits. Figures 8.6 and 8.7 present item characteristic curves for item 19 and item 9. Appendix G shows, in probability order, the location on the continuum, fit to the

measurement model and probability of fit to the model for the 30 items. Appendix E shows the thresholds for the 30 attitude items.

Table 8.1
Summary of fit statistics for the student attitude scale (30 items)

	Items	Students
Number	30	400
Location mean	0.00	0.99
Standard deviation	0.39	0.97
Fit statistic mean	0.15	-0.36
Fit statistic standard deviation	1.32	1.96
Item-trait interaction chi square = 165.40		
Degrees of freedom = 150		
Probability of item-trait (p) = 0.18		
Student Separation Index = 0.92		
Cronbach Alpha = 0.92		
Power of test-of- fit: excellent (based on the Separation index)		

Notes on Table 8.1

1. The mean of the 30 item difficulties is constrained to zero by the measurement model.
2. When the data fit the model, the fit statistics approximate a distribution with a mean near zero and a standard deviation near one. The item fit is good and student fit is satisfactory, but not an excellent fit. Item global fit is better than student global fit.
3. The item-trait interaction indicates the agreement displayed with all the items across all students from different locations on the scale (good for these data).This means that a unidimensional measure has been made.
4. The Student Separation Index is the proportion of observed student attitude variance considered true (in this scale, 92% and is very high).
5. Numbers are given to two decimal places because the errors are between. 0.06 and 0.08.

The Index of Separation (akin to traditional reliability) for the data of 30 items attitude scale with four categories is 0.92 (see Table 8.1). This means that the proportion of observed variance considered true is 92 % and that the measures are well separated in comparison to the errors.

The items are well targeted against the attitude measures (see Figures 8.1, 8.4 and Appendix G). That is, the range of item thresholds match the range of attitude measures of the students on the same scale. The item threshold values range from -2.04 logits (SE=0.07) to + 2.18 logits (SE=0.07) and the student measures range from -1.96

logits to +5.82 logits. There are only 38 students whose attitude measures are more than +2.18 logits and hence not 'matched' against an item threshold on the scale. These results indicate that a good measurement scale of attitude has been created, that the data are reliable and consistent, that the errors are small relation to the measures, and that the power of the tests-of-fit are excellent.

More Detailed Comments

Ordered Threshold and Response Categories

In order to determine threshold values, the RUMM 2010 program estimates the boundaries between each pair of adjacent response categories where there are odds of 1: 1 of answering in either category. For an item to fit the measurement model, the thresholds need to be ordered in line with the response categories. The threshold values are ordered from low to high for each of the 30 items indicating that the students have answered consistently and logically, in line with the response format used (see Figure 8.1).

Figure 8.1 is in logits, the log odds of answering the response categories positively. Student attitude measures are placed on the left hand side of the scale and item thresholds are placed on the right hand side scale. 11.1 refers to the threshold between the response categories 0 and 1 for item 11; 11.2 refers to the threshold between the response categories 1 and 2; 11.3 refers to the threshold between the response categories 2 and 3 for the same item. These thresholds are ordered: 11.1 is easiest (difficulty is –1.1 logits), 11.2 is harder (difficulty is +0.4 logits), and 11.3 is hardest (difficulty is +1.5 logits), in line with the ordering of the response categories. Other item thresholds are labeled similarly. Generally, the first threshold is towards the easy end of the scale (as expected), the second threshold is harder, and the third threshold is harder still (as expected). This supports the conceptual model of the response categories.

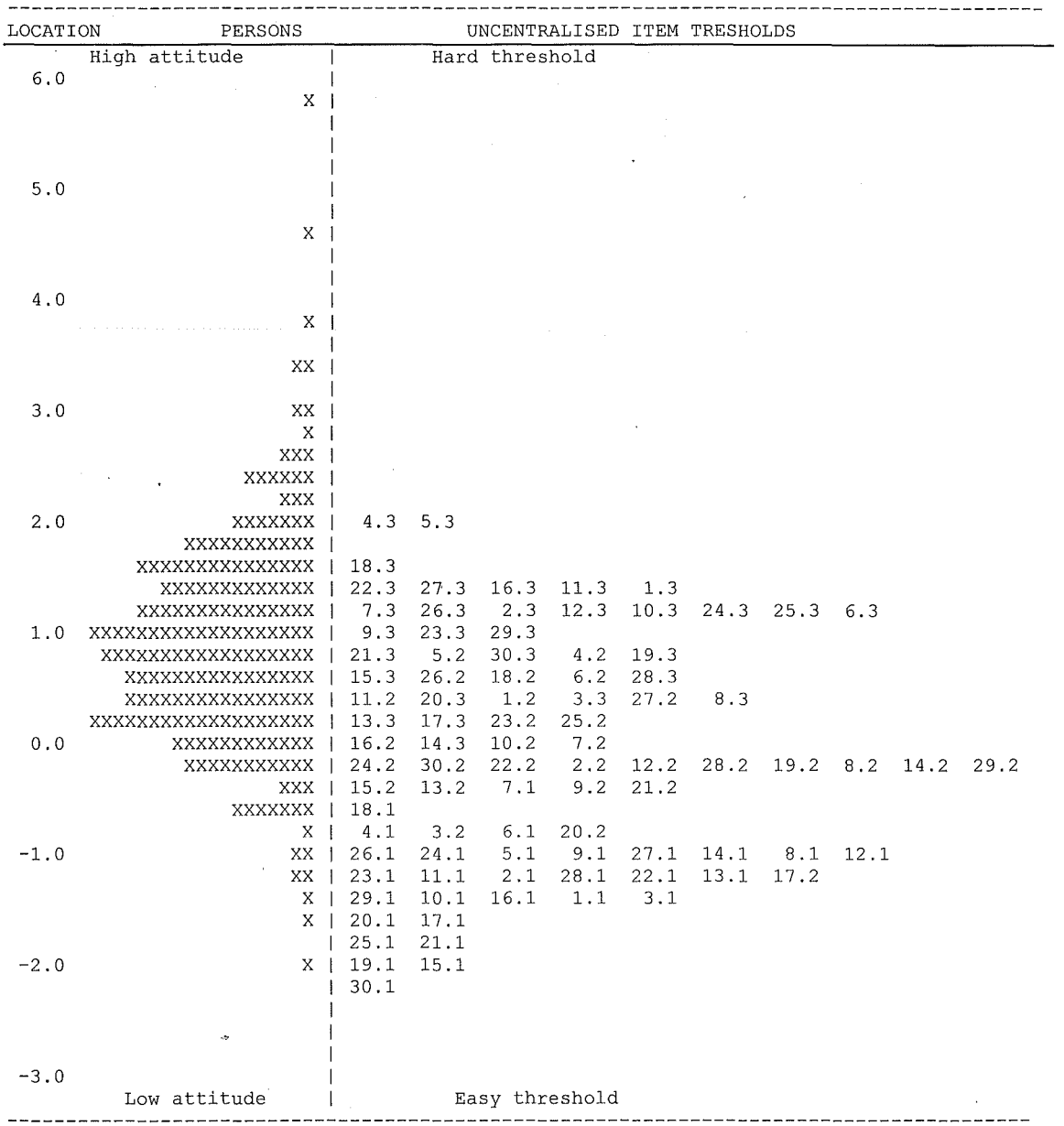


Figure 8.1 **Attitude measures and item thresholds (N=400, 4 categories, 3 thresholds for each of 30 items)**

Notes on Figure 8.1

1. The scale is in logits, the log odds of answering positively.
2. Measures of attitude are calibrated on the same scale as the item difficulties.
3. Measures are ordered from low to high on the left hand side and item thresholds are ordered from easy to hard on the right hand side.
4. Items at the easy end of the scale are answered positively by most students. As the items become harder, students need a higher attitude to answer the items positively.
5. Each x represents 2 students.
6. 1.1 = threshold 1 of item 1, 1.2 = threshold 2 of item 1, 1.3 = threshold 3 of item 1, and so on.

Table 8.2
An example of item thresholds for the attitude measure

No.	Item wording	Mean Threshold	Threshold 1	Threshold 2	Threshold 3
8.	I liked the Computerized Adaptive Test because of its immediate feedback.	-.107	-.838	-.046	.563
9.	I feel that it is worth having the chance to take the Computerized Adaptive Test.	-.056	-.898	-.313	1.042
10.	I feel like I am using my full ability with the Computerized Adaptive Test.	.011	-1.350	.029	1.352
11.	The Computerized Adaptive Test makes me want to study Mathematics.	.276	-1.145	.430	1.543

Note on Table 8.2

1. Thresholds are points between adjacent response categories where there is a 50% chance of answering in either category. Thresholds should be ordered in line with the ordering of the response categories for good measurement (good for these data and good for all items, see Appendix G).
2. For items 8, 9, 10, and 11, the thresholds (1, 2, 3) become harder in line with the ordering of the response categories: strongly disagree, disagree, agree, and strongly agree.
3. The mean threshold (item difficulty) becomes harder from item 8 to item 11 in line with the conceptual difficulties of items 8, 9, 10, and 11. Item 11 is conceptually more difficult than item 10, which is more difficult than item 9, which in term is more difficult than item 8.

Category Response Curves

The RUMM program provides a category response curve for each item, which makes it possible to view the ordering of the thresholds, and check whether the category responses are being answered logically and consistently. A perusal of the category response curves for the 30 items indicates that the students answered the response categories consistently and logically, resulting in ordered thresholds. For example, Figure 8.2 shows the category response curve for the good fitting item 19, *Computerized Adaptive Testing gives reliable results*. This is a positive item, where category 0 means strongly disagree, category 1 means disagree, category 2 means agree, and category 3 means strongly agree to the item wording.

Item 19 is a good-fitting item with a chi square probability of 0.99. Its difficulty is -0.32, indicating that students found it relatively easy to agree that the Computerized Adaptive Test gives reliable results for them. Figure 8.2 shows that the category curve 0 (category response strongly disagree) indicates that when a student has very low attitude (-6 logits), then the probability of answering in this category (strongly disagree) is 0.98 (very high as expected). As the student attitude increases (to -2 logits), then the probability of answering in this category drops to near 0.50 (as expected). When the

student attitude increases to+1 logits, then the probability of answering in this category drops to zero (as expected).

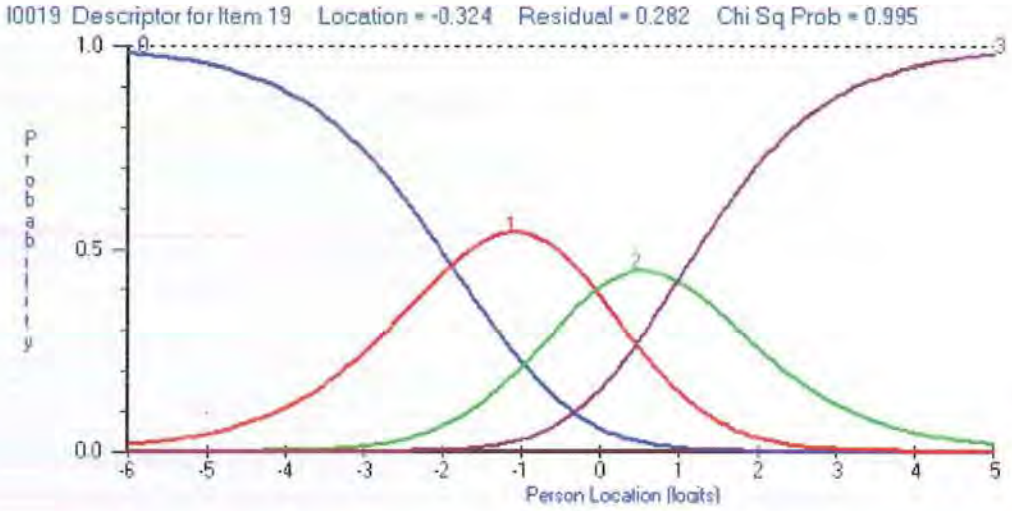


Figure 8.2 **Response category curve for item 19 (good-fitting item)**

Notes on figure 8.2

1. Threshold 1 is about -1.89 (boundary between category 0 and category 1).
2. Threshold 2 is about -0.57 (boundary between category 1 and category 2).
3. Threshold 3 is about +0.98 (boundary between category 2 and category 3).

For curve 1 (category response disagree), when the student has a very low attitude (-6 logits), then the probability of answering disagree is near zero (very low as expected). When the student attitude increases to -2 logits, then probability of answering disagree increases to 0.4 (as expected). When the student attitude increases to -1 logits, then the probability of answering disagree increases to 0.5 (as expected). When the student attitude increases to +3, the probability of answering disagree decreases to 0 (as expected).

For curve 2 (category response agree), when the student has a very low attitude (-3.5 logits), then the probability of answering agree is 0.0 (very low as expected). When the student attitude increases to -2 logits, then the probability of answering agree increases to 0.20 (as expected). When the student attitude increases to +0.5 logits, then the probability of answering agree increases to about +0.4 (as expected). When the student attitude increases to +5 logits, the probability of answering agree drops to zero (as expected).

For curve 3 (category response strongly agree), when the student has a very low attitude (-2 logits), then the probability of answering strongly agree is 0.0 (as expected). When the student attitude increases to 1.0 logits, then the probability of answering strongly agree increases to 0.40 (as expected). When the student attitude increases to +5 logits, the probability of answering strongly agree increases to 1.00 (as expected).

Item 9 is a medium difficulty item that doesn't fit the measurement model as well as one would like. Nevertheless, its thresholds are ordered and the Response Category Curve (see Figure 8.3) is acceptable. It has a moderate difficulty of -0.05 on this scale, which indicates students found it moderately easy to agree that *it is worth having the chance to take the Computerized Adaptive Test.* Figure 8.3 shows that the curve 0 (category response strongly disagree) indicates that when a student has a very low attitude (-4 logits), then the probability of answering strongly disagree is 0.95 (very high as expected). As the student attitude increases to -2 logits, then the probability of answering strongly disagree drops to 0.70 (as expected). When the student attitude increases to +1 logits, then the probability of answering strongly disagree drops to zero (as expected).

For curve 1 (category response disagree), when the student has a very low attitude (-4 logits), then the probability of answering disagree is 0.05 (very low as expected). When the student attitude increases to -2 logits, then probability of answering disagree increases to 0.5 (as expected). When the student attitude increases to -1 logits, then the probability of answering disagree increases to 0.4 (as expected). When the student attitude increases to +3 logits, then the probability of answering disagree decreases to 0 (as expected).

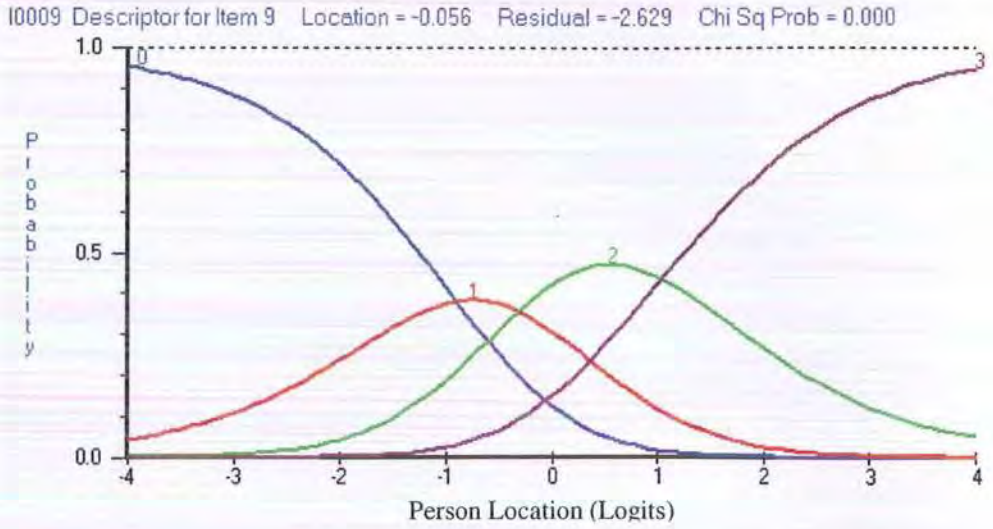


Figure 8.3 **Response category curve for item 9 (not-so-good fitting item)**

Notes on figure 8.3

1. Threshold 1 is about -.89 logits (boundary between category 0 and category 1).
2. Threshold 2 is about -.31 logits (boundary between category 1 and category 2).
3. Threshold 3 is about +1.04 logits (boundary between category 2 and category 3).
4. While the thresholds are separated, the response category curves are not as well separated as would be liked and this may be partly contributing o the poor fit of the item to the measurement model.

For curve 2 (category response agree), when the student has a very low attitude (-3 logits), then the probability of answering agree is 0.0 (very low as expected). When the student attitude increases to -2 logits, then the probability of answering agree increases to 0.04 (as expected). When the student attitude increases to +1 logits, then the probability of answering agree increases to 0.45 (as expected). When the student attitude increases to +5 logits, then the probability of answering agree drops to zero (as expected).

For curve 3 (category response strongly agree), when the student has a very low attitude (-2 logits), then the probability of answering strongly agree is 0.0 (very low as expected). When the student attitude increases to +1 logits, then the probability of answering strongly agree increases to 0.45 (as expected). When the student attitude increases to +4 logits, then the probability of answering strongly agree increases to 0.95 (as expected).

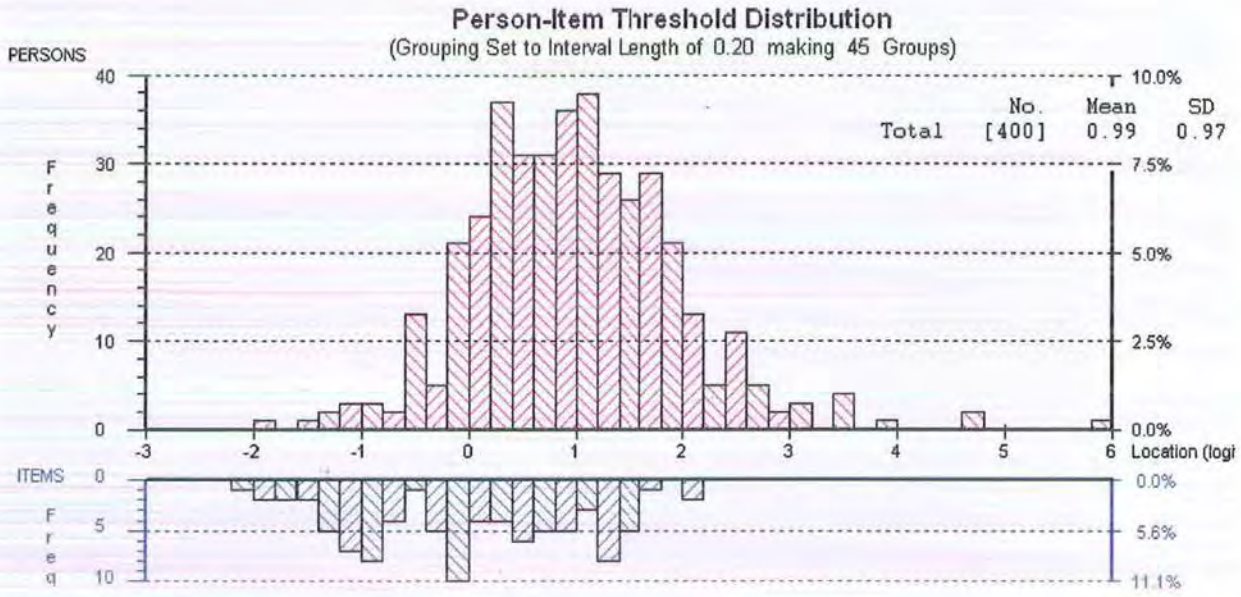


Figure 8.4 **Item thresholds and attitude measures on the same scale**

Note on figure 8.4

1. There are three thresholds per item, corresponding to the odds (1:1) of answering in the adjacent response categories. The thresholds are ordered in line with the ordering of the response categories from low to high.
2. Thresholds are ordered from easy to hard on the lower side of the scale in logits.
3. Student measures are ordered from low to high on the upper side of the scale in logits.

The locations (difficulties) of the items cover the lower and the middle ranges of attitude measures, but not the higher range as well as it could (see Figure 8.5). However, the thresholds of the items range from about -2.0 to +2.0 logits and cover more of the range of attitude measures (about –1.9 to +5.8 logits, see Figure 8.4). This means that, while the targeting of the attitude items is acceptable, harder items could be added in a revision of the scale, to cover the higher measures (+2.0 to +6.0 logits).

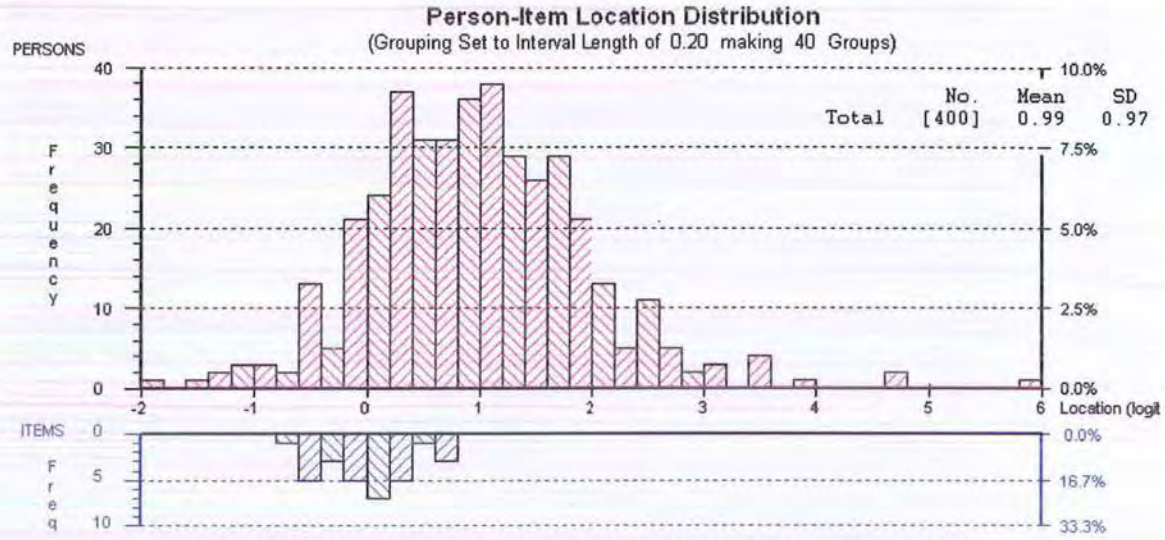


Figure 8.5 **Attitude measures and item locations on the same scale**

Notes on figure 8.5

1. The scale is in logits, the log odds of answering the response categories.
2. Attitude measures (low to high) are placed on the upper side of the scale and item locations (difficulties) from easy to hard are placed on the lower side of the scale.

Attitude measures are displayed on the same scale as the item difficulties in Figure 8.5 from –2.0 to zero logits, there are 51 (12.8%) students and these can be regarded as having a low attitude. From zero to +1.0 logits, there are 159 (39.8%) students and these can be regarded as having a moderately high attitude (they can answer some items with a high response category). From +1.0 to +3.0 logits, there are 179 (44.8%) students and they can be regarded as having a high attitude. From +3.0 to +6.0 logits, there are 11 (2.8%) students and they can be regarded as having a very high attitude in using the attitude questionnaire.

Item Characteristic Curves

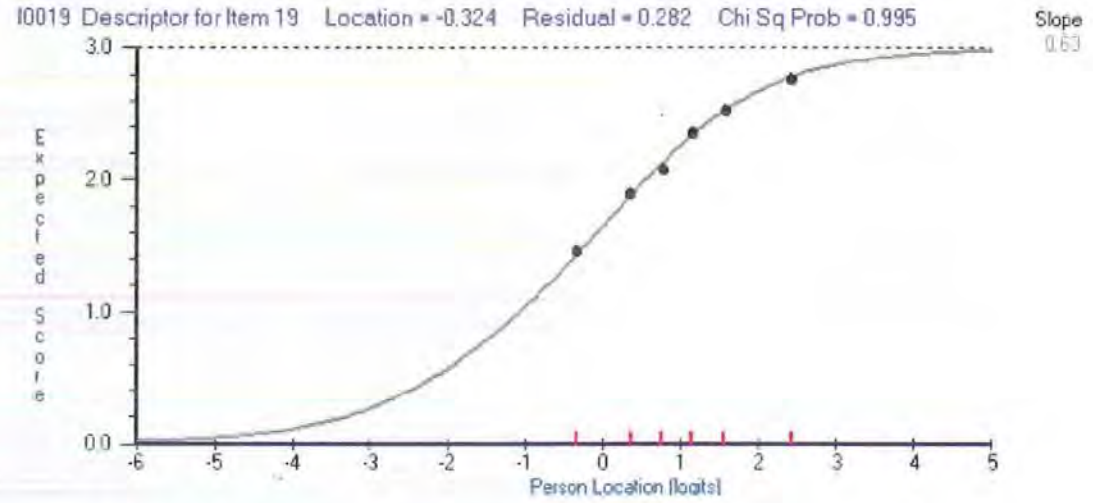


Figure 8.6 **Characteristic curve for item 19**

The item characteristic curve for Item 19 (good fitting item) of the attitude scale is shown on Figure 8.6. The line indicates the expected score of attitude groups, ranging from the lowest to highest attitude groups. Each black dot represents the observed score of a student attitude group. When the observed scores closely follow the curve of expected values, the group is performing as expected on the item. Item 19 shows good fitting item to the model with all groups attitude close to the expected scores.

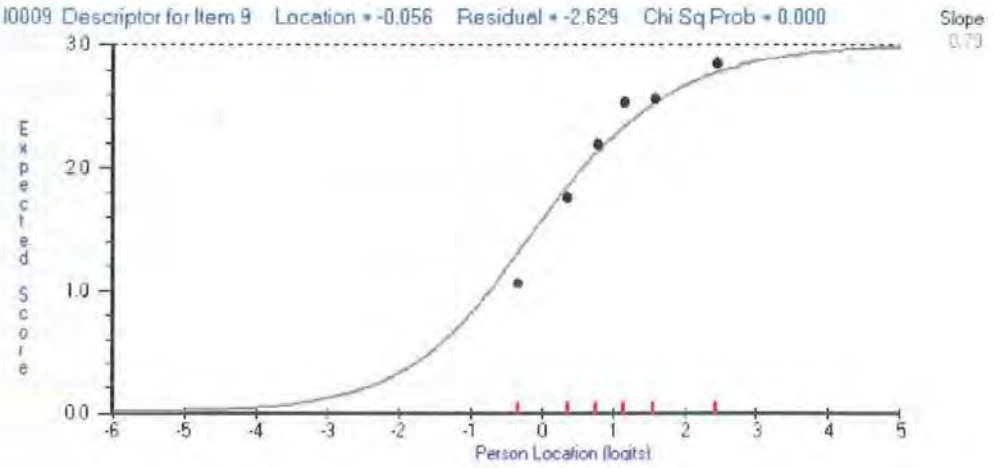


Figure 8.7 **Characteristic curve for item 9**

Item 9 is a not so good fitting item of the attitude scale. The item characteristic curve is shown on Figure 8.7. The one lower attitude group has performed slightly lower than expected and the one higher attitude group has performed slightly higher than expected, on this item, whereas the others two medium and two higher attitude groups are performing as expected. This is demonstrated by the item characteristic

curve for item 9 where the black spot representing the one lower attitude group appears below the black line (which represents the expected score), the black spot representing the one higher attitude group appears above the black line, and the black spots representing the two medium and two higher attitude groups appear close to the black line. Item 9 shows a not so good fitting item to the model because there are 2 from 6 attitude groups (33.33%) not close to the expected measures.

Item Difficulties

After the Rasch analysis, the items were ordered in terms of their calibrated item difficulties (see Tables 8.3, 8.4, 8.5, 8.6, and 8.7) by sub-groups.

Table 8.3
Item difficulties in order for Like and Interest in CAT (N=400)

Item No.	Item wording	Difficulty
	Like and Interest in CAT	
1. (3)	The Computerized Adaptive Testing is very interesting.	-0.53
2. (14)	I am happy doing the Computerized Adaptive Test without limited time.	-0.29
3. (8)	I like the Computerized Adaptive Test because of its immediate feedback.	-0.11
4. (2)	I am happy and enjoyed doing a Computerized Adaptive Test.	+0.03
5. (12)	I feel lucky to have the chance to take a Computerized Adaptive Test.	+0.14
6. (1)	I am enthusiastic about taking part in a Computerized Adaptive Test.	+0.23
7. (6)	I liked the Computerized Adaptive Test because it was not too difficult for me.	+0.48

Notes on table 8.3
1. Item difficulties are in logits.
2. The difficulties are reported to 2 decimal places because the errors are between 0.06 and 0.08.
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The students found it very easy to say that Computerized Adaptive Testing is very interesting (item 3) and they are happy with it (item 14). They found it easy (but harder) to say that they like it because of its immediate feedback (item 8). They found it moderately easy (but harder) to say that they enjoyed doing the Computerized Adaptive Test (item 2) and moderately hard to say that they are lucky to have the chance to take a Computerized Adaptive Test (item 12). They found it harder to say that they were enthusiastic about taking part in a Computerized Adaptive Test (item 1) and this was expected because the attitude is linked to a behaviour which is theoretically harder than a similar attitude (compare item 3 and item 8). They found it very hard to say that they liked the Computerized Adaptive Test because it was not too difficult (item 6) because this again is linked to a behaviour (compare item 3 and item 8).

Table 8.4
Item difficulties in order for confidence with and Use of CAT (N=400)

Item No.	Item wording	Difficulty
	Confidence with and Use of CAT	
1. (13)	I want Computerized Adaptive Testing to be used for my other subjects.	-0.41
2. (9)	It is worth taking a Computerized Adaptive Test.	-0.06
3. (10)	I feel like using my full ability with the Computerized Adaptive Test.	+0.01
4. (11)	The Computerized Adaptive Testing makes me want to study Mathematics.	+0.28
5. (7)	After finishing the Computerized Adaptive Testing, I feel like wanting to do another.	+0.37
6. (5)	I believe that I can do the Computerized Adaptive Test well.	+0.72
7. (4)	I took the Computerized Adaptive Test with confidence.	+0.75

Notes on table 8.4
1. Item difficulties are in logits.
2. Item 6(5) and 7(4) difficulties are equal within the measurement error (0.06 to 0.08).
3. Items are ordered from easy (top) to hard (bottom).
4. Original item numbers given in brackets.

The students found it very easy to say that they want Computerized Adaptive Testing for all other subjects (item 13) and much harder (but still moderately easy) to say that they feel it is worth taking a Computerized Adaptive Test (item 9). They found it moderately easy to say that they feel like using their full ability with the Computerized Adaptive Test (item 10) and much harder to say that Computerized Adaptive Testing makes them want to study Mathematics (item 11). They found it hard to say that after finishing the Computerized Adaptive Testing, they want to do another (item 7), in line with their answer to item 11. They found it extremely hard to say that they believe they can do the Computerized Adaptive Test well (item 5) and that they took the Computerized Adaptive Test with confidence (item 4), as would be expected.

Table 8.5
Item difficulties in order for CAT as Modern and Useful (N=400)

Item No.	Item wording	Difficulty
	CAT as Modern and Useful	
1. (17)	Computerized Adaptive Testing is modern.	-0.73
2. (20)	The Computerized Adaptive Testing is currently appropriate for these days.	-0.56
3. (15)	Computerized Adaptive Testing is very useful.	-0.53
4. (23)	Computerized Adaptive Testing allows students to spend less time on testing.	+0.07
5. (24)	Computerized Adaptive Testing provides examinees with appropriate items.	+0.08
6. (18)	Computerized Adaptive Testing saves money.	+0.61

Notes on table 8.5

1. Item difficulties are in logits.
2. The difficulties of item 2 (20) and 3 (15) are equal within the measurement error (0.06 to 0.08).
3. The difficulties of item 4 (23) and 5 (24) are equal within the measurement error (0.06 to 0.08).
4. Items are ordered from easy (top) to hard (bottom).
5. Original item numbers given in brackets.

The students found it very easy to say that Computerized Adaptive Testing is modern (item 17) and a little easier to say that it is appropriate for these days (item 20), and useful (item 15). They found it moderately easy to say that Computerized Adaptive Testing allows students to spend less time on testing (item 23) (with the implication that

they can then spend more time on learning) and that it provides students with appropriate items (item 24). Students found very hard to say that Computerized Adaptive Testing saves money (item 18).

Table 8.6
Item difficulties in order for CAT as Reliable, Fair and Good (N=400)

Item No.	Item wording	Difficulty
	CAT as Reliable, Fair and Good	
1. (21)	Computerized Adaptive Testing is fair for all students.	-0.36
2. (19)	Computerized Adaptive Testing gives reliable results.	-0.32
3. (25)	Computerized Adaptive Testing makes examinees careful when doing the test.	-0.01
4. (16)	Computerized Adaptive Testing is challenging.	+0.05
5. (22)	Computerized Adaptive Testing inspires the students to do the test.	+0.08

Notes on table 8.6

1. Item difficulties are in logits.
2. The difficulties of item 1 (21) and 2 (19) are equal within the error of measurement (0.06 to 0.08).
3. The difficulties of item 3 (25), 4(16) and 5 (22) are equal within the error of measurement (0.06 to 0.08).
4. Items are ordered from easy (top) to hard (bottom).
5. Original item numbers given in brackets.

Students found it very easy to say that Computerized Adaptive Testing is fair for all students (item 21) and that it gives reliable results (item 19). They found it moderately easy to say that Computerized Adaptive Testing makes student take care in testing (item 25), provides a challenge (item 16), and inspires the students to do the test (item 22).

Table 8.7
Item difficulties in order for CAT Recommendations (N=400)

Item No.	Item wording	Difficulty
	CAT Recommendations	
1. (30)	I am ready to apply the knowledge from Computerized Adaptive Testing.	-0.43
2. (28)	If possible, I 'd rather take a Computerized Adaptive Test.	-0.13
3. (29)	If I have a chance, I will introduce my younger friends to Computerized Adaptive Testing.	-0.11
4. (26)	I wish I could take a Computerized Adaptive Test in a Mathematics test competition.	+0.33
5. (27)	I will tell my friends about Computerized Adaptive Testing.	+0.36

Notes on table 8.7

1. Item difficulties are in logits.
2. Difficulties for item 2 (28) and 3 (29) are equal within the error of measurement.
3. Difficulties for item 4 (26) and 5 (27) are equal within the error of measurement.
4. Items are ordered from easy (top) to hard (bottom).
5. Original item numbers given in brackets.

Students found it very easy to say that they were ready to apply their knowledge of Computerized Adaptive Testing (item 30). They found it moderately easy to say that they would rather take a Computerized Adaptive Test (item 28) (than an ordinary test) and that they would introduce their younger friends to Computerized Adaptive Testing (item 29). Students found it very hard to say that they could take a Computerized Adaptive Test in a mathematics competition (item 26) and that they would tell their friends about Computerized Adaptive Testing (item 27).

Summary

The computer program RUMM (Andrich et al., 2003) was very useful in analyzing data on student attitude towards a Computerized Adaptive Testing. The Rasch analysis showed that:

1. Twenty-seven of the 30 items fitted the measurement model with a probability >0.04;

2. There was good global item fit to the measurement model;

3. Global person fit to the measurement model was not as good as the global item fit, but was acceptable;

4. The item-trait interaction chi-square was not statistically significant, indicating that a uni-dimensional trait (or at least a dominant trait) was measured;

5. The Student Separation Index was 0.92, indicating that the errors were small in relation to the separation of measures along the scale;

6. The thresholds for the 30 items were ordered in line with the ordering of the response categories, meaning that the students used the response categories logically and consistently;

7. The targeting of the item thresholds against the student measures was reasonably good but, in any revision of the scale for these students, some harder thresholds (items) have to be added to cater for those with higher attitudes and behaviour towards Computerized Adaptive Testing.

This evidence shows that a reliable scale was constructed from which valid inferences and conclusions could be drawn.

The scale of attitude towards Computerized Adaptive Testing (CAT) was shown to consist of:

1. Seven items relating to Like and Interest in CAT, ordered from very easy (CAT is very interesting, item 13) to very hard (I liked CAT because it was not too difficult for me, item 6);

2. Seven items relating to Confidence with and Use of CAT, ordered from very easy (I want a CAT to be used for other subjects, item 13) to extremely hard (I took CAT with confidence, item 4) ;

3. Six items relating to CAT as Modern and Useful, ordered from extremely easy (CAT is modern, item 17) to extremely hard (CAT saves money, item 18);

4. Five items relating to CAT as Reliable, Fair and Good, ordered from very easy (CAT is fair for all students, item 21) to moderately hard (CAT inspires students to do the test, item 22);

5. Five items relating to CAT Recommendations, ordered from very easy (I am ready to apply knowledge from CAT, item 30) to very hard (I will tell my friends about CAT, item 27).

Some valid and important inferences that can be drawn from the linear scale of attitudes to Computerized Adaptive Testing are now listed.

Students had a very positive attitude towards Computerized Adaptive Testing.

(a) Ninety-six percent of students (384/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 20, Computerized Adaptive Testing was appropriate.

(b) Ninety-six percent of students (384/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 3, Computerized Adaptive Testing was interesting.

(c) Ninety-one percent of students (364/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 15, Computerized Adaptive Testing was very useful.

(d) Ninety-one percent of students (364/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 21, Computerized Adaptive Testing was fair to all students.

(e) Ninety-one percent of students (364/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 13, Computerized Adaptive Testing should be used for all their other subjects.

(f) Eighty-five percent of students (342/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 19, Computerized Adaptive Testing gave reliable results.

(g) Eighty-five percent of students (342/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 28, I would rather take a Computerized Adaptive Test (than an ordinary classroom test).

Students indicated that there was some apprehension about taking the Computerized Adaptive Test, probably in part, because it was new to the students.

(a) Fifty-nine percent of students (236/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 5, indicating that they believed that they could do the Computerized Adaptive Test well.

(b) Fifty-nine percent of students (236/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 5, indicating that they believed that they took the Computerized Adaptive Test with confidence.

The five attitudes that students found most hard were indicated by the five highest, third-level thresholds (agree/strongly agree) and they are listed in order from hard to easier.

(a) I took the Computerized Adaptive Test with confidence (item 4) (hardest).

(b) I believe that I can do the Computerized Adaptive Test well (item 5).

(c) Computerized Adaptive Testing saves money (item 18).

(d) Computerized Adaptive Testing inspires students to do the test (item 22).

(e) I will tell my friends about Computerized Adaptive Testing (item 27).

The five attitudes that students found the five easiest to hold were indicated by the lowest, first-level thresholds (strongly disagree/disagree) and they are listed in order from easiest to harder.

(a) I am ready to apply the knowledge from Computerized Adaptive Testing (item 30) (easiest).

(b) Computerized Adaptive Testing gives reliable results (item 19).

(c) Computerized Adaptive Testing is very useful (item 15).

(d) Computerized Adaptive Testing makes examinees careful when doing the test (item 25).

(e) Computerized Adaptive Testing is fair for all students (item 21).

The next chapter (Chapter Nine) answers the research questions, and explains the implications of the study for students and teachers, for schools and administrators, for future research and comments on the non-fitting items.

CHAPTER 9

RESEARCH QUESTIONS AND IMPLICATIONS

In this chapter the research questions of the study are answered. Pedagogical implications are discussed and suggestions offered for the implementation of the findings and recommendations, and for further research.

Research Questions

The research questions can now be answered:

Research question 1: *Can the difficulties of the items in the 'bank' be modelled and aligned on a scale of Mathematics achievement from easy to hard using a Rasch measurement model*?

The answer to the first research question is that the difficulties of the items in the bank can be modelled and aligned on a scale of Mathematics achievement from easy to hard using a Rasch measurement model. Ninety-eight of the 290 items fitted the measurement model. There were seven items relating to the identification of an equation, 11 items relating to the identification of the true equation, three items on identifying equations with an unknown, eight items on finding the value of an unknown that satisfies the equation, 17 items relating to Identify the Method to solve the Equation, 12 items relating to finding the solutions to equations, 23 items relating to finding a solution of an equation which related the given condition, ten items on selecting an equation converted from a verbal problem or a verbal problem related to an equation, and seven items on problem solving. Their difficulties were calibrated on the same scale together with student measures of mathematics so that the ordering of students with high, medium, and low measures is in accordance with the difficulties of the items.

Research question 2: *Can the Mathematics Computerized Adaptive Testing software be used to examine differences in mathematics competency of Year 6 students in Thailand?*

The answer to the second research question is that the Mathematics Computerized Adaptive Testing software can be used to examine differences in mathematics competency of Year 6 students in Thailand. The findings indicated that the mathematics competencies were significantly different at p=0.05 among four groups of stopping criteria (F = 5.09, df = 3, 396, p = 0.00). The mean highest mathematics competency (+0.63 logits) and the mean lowest mathematics competency (+0.38 logits) were in group 4 (stopping criteria is $SEE_m - SEE_{m-1} \leq 0.005$) and group 2 (stopping criteria is $SEE \leq 0.30$).

Research question 3: *Are changes in test length and testing times related to different stopping criteria in Computerized Adaptive Testing?*

The answer to the third research question is that there are changes in test length and testing times related to differences in stopping criteria in Computerized Adaptive Testing. Test lengths and testing times were significantly different at p=0.05 among four groups of stopping criteria (F = 191.30, df = 3, 396, p = 0.00;. F = 53.85, df = 3, 396, p = 0.00; and F = 5.09, df = 3, 396, p = 0.00). The mean highest test length and testing times (8.14 items and 5.74 minutes) and the mean lowest test length and testing times (3.14 items and 2.38 minutes) were in group 3 (stopping criteria is $SEE \leq 0.40$) and group 2 (stopping criteria is $SEE \leq 0.30$). Each group was significantly different in both test length and testing times at p=0.05 from the others.

Research question 4 : *Are changes in test length and testing times related to differences in mathematics competency of the examinees?*

The answer to the fourth research question is that there are changes in test length and testing times related to difference in mathematics competency of the examinees. Test lengths and testing times were significantly different at p=0.05 among the three groups of mathematics competencies (F = 16.91, df = 2, 397, p = 0.00 and F = 4.07, df = 2, 397, p = 0.02). The mean highest test length and testing times (7.07 items and 4.92 minutes) and the mean lowest test length and testing times (5.21 items and 3.97 minutes) were in group 1 (low mathematics competency) and group 2 (moderately high mathematics competency). There was an only one significantly different test length and

also testing times at p=0.05 between students in group1 (low mathematics competency) and group 2 (moderately high mathematics competency).

Research question 5: *Can the attitude to the Mathematics Computerized Adaptive Testing of the Year 6 students in Thailand be measured using a Rasch measurement model and aligned from low to high on the same scale?*

The answer to the fifth research question is that the attitude to the Mathematics Computerized Adaptive Testing of the Year 6 students in Thailand can be measured using a Rasch measurement model. Attitude measures and attitude item difficulties were calibrated together on the same linear scale. The findings indicated that 27 of the 30 items fitted the measurement model with a probability >0.04. The student measures (N = 610) and the item difficulties (I=30) were calibrated on the same linear scale where a uni-dimensional (or dominant) trait influenced all the items. The thresholds for the 30 items were ordered in line with the ordering of the response categories, meaning that the students used the response categories logically and consistently. While the data for the 30 items were reliable, some harder items have to be added to cater for those with the highest attitudes towards computerized adaptive testing and some easier items have to be added to cater for those with the lowest attitudes towards computerized adaptive testing.

Research question 6: *What are Thailand Year 6 student abilities in mathematics and attitudes towards the Mathematics Computerized Adaptive Testing?*

For the mathematics ability, the answers to the research question is that there were 72.25 %, 16.75%, and 8% of the Prathom Suksa 6 students having a moderately high, low, and high mathematics achievement respectively.

For the attitudes towards the Mathematics Computerized Adaptive Testing, the findings indicated that students had a very positive attitude towards the computerized adaptive testing. Ninety-six percent of students (384/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 20, computerized adaptive testing was appropriate and item 3, computerized adaptive testing was interesting. Ninety-one percent of students (364/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 15, computerized adaptive testing was very useful, item 21, computerized adaptive testing was fair to all students and item 13,

computerized adaptive testing should be used for all their other subjects. Eighty-five percent of students (342/400) had a measure equal to, or greater than, the second threshold (disagree/agree) of item 19, computerized adaptive testing gave reliable results and item28, I would rather take a computerized adaptive test (than an ordinary classroom test).

Research question 7: *Are there changes in measured mathematics ability using Computerized Adaptive Testing when different stopping criteria are applied?*

The answer to the research question is that measured mathematics competencies were significantly different among four groups of stopping criteria. The mean highest and the mean lowest mathematics competencies were in group 4 (stopping criteria is $SEE_m - SEE_{m-1} \leq 0.005$) and group 2 (stopping criteria is $SEE \leq 0.30$).

Implications

This part presents the implications for those who are involved in the assessment of students' learning achievement such as students, teachers, schools, and school administrators. In addition, suggestions for further research are also given.

For Students and Teachers

With regard to teachers, computerized adaptive testing is likely to be accurate in assessing individual student's ability in any tested situation. The teachers can use it with individual students or groups without worrying about cheating in the examinations. Computerized Adaptive Testing could help prevent examinees from getting bored with having too many test items. Also, through Computerized Adaptive Testing, each examinee does different test items and different number of items. This depends upon an individual's ability. In addition, data gained from the test can be used for many purposes, such as, to follow up an individual's learning progress, to diagnose deficiencies in each student, and to assess students' achievement. Student's weaknesses in any subject matter can consequently be remedied. Computerized Adaptive Testing is an efficient and authentic assessment of student's learning. It is recommended that teachers prepare more examples of item banks and Computerized Adaptive Testing for use in primary schools in Thailand in different subjects.

For Schools and Schools Administrators

In relation to school network, it would be useful for members of the network to access the item banks available through Computerized Adaptive Testing. The school network could either develop a bank containing tests of different subject areas or different banks for different subject areas. This can be done by establishing one school as the item bank, equipped with a central computer, while other member schools in the network can access the bank through the networking computers in their schools. This can save time and school resources in preparing tests and conducting examinations whenever they need. Regarding the development of the test items, teachers in every school network could cooperate to construct, try out, analyse, and select qualified items to store the item bank. If this process is continuously done, the item bank will become large with thousands of well-calibrated items by difficulty equated on the same scale. The pooling of resources between different schools might be launched by provincial administrators. The provincial administrators could run in-service courses on CAT and item banks, with items appropriate to many school subjects. Moreover, Computerized Adaptive Testing is a new approach for learning assessment and evaluation which is likely to be the future of assessment. There is a large monetary cost to implement this, but it would be well worth it, and probably necessary in the future.

For Future Research

For the future research, there are several recommendations:

In creating an item bank with CAT, a large number of items of different difficulties are required so that the item difficulties cover a continuum of examinees' ability, ranging from very low to very high ability. This will assure the accuracy of a measurement and suit examinees at all ability levels. The non-fitting items in the present study should be re-worded, new data collected, and the analysis repeated (more on this late in this chapter).

In response to the finding that harder items as well as easier items need to be included to the item bank so as to better target the high achievers and the low achievers (see also pages 88 and 90), it is suggested that in any future research a Distractor Analysis should be conducted using the latest RUMM 2020 program (Andrich, Sheridan, and Luo, 2005). Distractors are the alternatives which are generally scored 0, while the correct answer is scored 1. Because distractors are not likely to be equally incorrect, methods which would increase the precision of measurement by identifying

information in distractors are useful and Distractor Analysis is one of them. When a distractor is very easy, the possibility of guesswork increases. This in turn affects the weightings (and subsequently the logit values) of the items. By conducting a distractor analysis, one can detect how the distractors can be improved to reduce guessing and thereby improve targeting.

To attract examinees' attention, concentration and willingness, a test in a form of multimedia, i.e., with moving graphics, pictures and sound should be constructed. This will make the test items more challenging.

At present, the Internet is becoming more significant in all subject areas. For instance, in giving presentations, in teaching and learning, in conducting research, in transferring information, and in measurement and evaluation. In the area of measurement and evaluation, further research on conducting CAT on the internet is suggested.

Generally, most of the CATs have been created on cognitive traits. Only a few have been constructed on affective traits, i.e., diligence, discipline, honesty, attitudes. Such traits have been measured using rating scale questionnaire. There are only a few qualified instruments for measuring the affective traits in Thailand. Therefore, more studies on CAT for the affective traits employing the Partial Credit Model of Rasch in analyzing the questionnaire items are worth conducting.

For Items Not Fitting the Rasch Measurement Model in the Present Study

There were 182 items in the present study that did not fit the measurement model and were deleted. This raises the question: "Why is this so and what can be done about it?". The RUMM2010 program produces a great of output relating to fit to the measurement model, but it doesn't tell the researcher why any particular items do not fit, only that they do not fit. All the non-fitting items in the present study were re-examined with a view to working out what might be wrong. It appeared that there might be two main reasons for the non-fit.

Reason One: It is likely that the wording of some items confused some students. The test items were written in English (for the supervisor and author) and then translated into Thai (for student use in the schools). It would seem that this caused some expression problems which resulted in some different item interpretations by some students. This, in turn, resulted in some disagreements by some students in relation to item difficulties which caused the misfit. If students cannot agree about the item

difficulties in Rasch measurement, then the Rasch output shows this. It is suggested that the non-fitting items be re-worded, re-implemented and re-analyzed.
Reason Two: It may be that some students treated the mathematics test just like school work and didn't take enough care in answering some items, thus making some careless mistakes. This, in turn, caused some disagreement amongst students about item difficulties and thus misfit to the model.

It should be noted that, if more items fit the measurement model after re-wording and re-analysis, then the final item bank will be improved considerably. The clear implication from this study is that item banking with multi-media items and Computer Assisted Technology is the assessment future for schools.

REFERENCES

Anderson, E. (1995). Polytomous Rasch models and their estimation. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 271-291). New York: Springer-Verlag.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 30*, 84-98.

Andrich, D. (1982). Using latent trait measurement model to analyse addtitudinal data: A synthesis of viewpoints. In D. Spearritt (Ed.), *The improement of measurement in educational and psychology :Contributions of latent trait theiries* (pp. 89-126). Melbourne: Australian Council for Educational Research.

Andrich, D. (1988a). A general form of Rasch's extended logistic model ofor partail cradit scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D. (1988b). *Rasch models for measurement.* Newbury Park, California: Sage Publications.

Andrich, D. (1988a). A general form of Rasch's extended logistic model ofor partail cradit scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D. (1988b). *Rasch models for measurement.* Paper presented at the Sage university paper on quantitative applications in the social sciences,series number 07/068, Newbury Park, California.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J.A.Keats, R.Taft, R.A.Health & S.Lovibond (Eds.), *Mathematical and theotical systems* (pp. 7-16). North Holland, Amsterdam: Elsevier Science Publishers.

Andrich, D., Sheridan, B., & Luo, G. (2003). *RUMM2010:A windows-based item analysis program employing Rasch unidimensional measurement models.* Perth, Western Australia: RUMM Laboratory.

Andrich, D., Sheridan, B., & Luo, G. (2005). *RUMM : A windows-based item analysis program employing Rasch unidimensional measurement models.* Perth: Murdoch university.

Andrich, D., & van Schoubroeck, L. (1989). The general health questionnaire: a psychometric analysis using latent trait theory. *Psychological medicine, 19*, 469-485.

Baghi, H., Ferrara, S., & Gabrys, R. (1992, April). *Student Attitudes toward Computer-Adaptive Test Administrations.* Paper presented at the The annual meeting of the American Educational Research Association,, San Francisco, CA.

Baghi, H., Gabrys, R., & Ferrara, S. (1991, April). *Applications of computer-adaptive testing in Maryland.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Beevers, C. E., McGuire, G. R., Stirling, G., & Wild, D. G. (1995). Mathematical ability assessed by computer. *Computers & Education, 25*(3), 123-132.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models and their uses inferring an examinee's abilities. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categoris. *Psychometrika, 37*, 29-51.

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

Bock, R. D., & Mislevy, R. J. (1982a). Adaptive EAP estimation of ability in a micro computer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Bock, R. D., & Mislevy, R. J. (1982b). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725-737.

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

Boonprasert, U. (1988). *The construction of item banking*. Bangkok: Chulalongkorn University.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). *The foue generation of computerized educational measurement* (No. ED 395 000). Princeton, NJ: Educational Testing Service.

Cavana, R. Y., Delahaye, B. L., & Sekaran, U. (2001). *Applied business research :qualitative and quantitative methods*. Singapore: Markono Print Media Pte Ltd.

Chansilp, S. (2006). *Online Test Bank SUT*. Retrieved 20/5/2007, from http://library.sut.ac.th/central/HeaderFrame.html

Choppin, B. (1981). Educational measurement and the item bank model. In C. Lacey & D. Lawton (Eds.), *Issues in evaluation and accoutability*. Methuen, London.

Choppin, B. (1985). Principles of item banking. *Evaluation in Education, 9*, 87-90.

Cordova, C., & Mario, J. (1998). *Applications of network flows to computerized adaptive testing (Item Response, test assembly).* Unpublished PhD, Rutgers The State University of New Jersey, NJ.

Department of Academics. (1991). *Local item bank for schools [Thai Publication. Translation of Title]*. Bangkok: Khurusapha press.

Douglas, G. (1982). Conditional inference in generic Rasch model. In D.Spearritt (Ed.), *The improvment of measurement in educational and psychology: contributions of latent traits theories*. Hawthorn, Victoria: Australian Council of Educational Research.

Dubeis, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement, 35*, 869-884.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*(5), 379-393.

Elliot, C. D. (1990). *Differentail abilities scales*. San Antonio, TX: The Psychological Corporation.

Embreston, S. E., & Reise, S. P. (2000). *Item response theory for phychologists*. Mahwah, NJ: Erlbaum.

Embretson, S. E., & Hershberger, S. L. (1999). Preface. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Gershon, R. C. (1990). *Rasch-model procedures used to build the JOCRF vocabulary item bank* (Technical report No. 1990-3). Chicago, IL: Johnson O' Connor Research Foundation.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

Gershon, R. C., & Bergstrom, B. A. (Artist). (1995). *Does cheating on CAT pay: Not!* [ERIC Document Reproductions No. TM024692].

Glastonbury, B., & MacKean, J. (1991). Surveys methods. In G. Allan & C. Skinner (Eds.), *Handbook for research students in the social sciences* (pp. 225-247). London: Falmer Press.

Green, B. F. (1984). Technical guidlines for assessing computerized adaptive test. *Journal of Educational Measurement, 21*, 72, 97, 347,352.

Gronlund, N. E. (1998). *Assessment of student achievement* (6 ed.). Boston: Allyn& Bacon.

Hambleton, R. K. (1986). The changing conception of measurement : A commentary. *Applied Psychological Measurement, 10*, 415-421.

Hambleton, R. K., Sawaminathan, H., & Rogers, J. H. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory : Principles and application*. Boston: Kluwer-Nijhoff Publishing.

Henly, S. J., Klebe, k. J., McBride, J. R., & Cudeck, R. (1989). Adaptive and conventional versions of the DAT: The first complete test bettery comparison. *Applied Psychological Measurement, 13*, 363-371.

Hiscox, M. D. (1983). *A balance sheet for educational item banking.* Paper presented at the annual meeting of National Council for Measurement in Education, Montreal.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*, 16-25.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: application to psychological measurement*. Homewood, IL: Dow-Jones Irwin.

Jones, D. H. (1982). Redescending M-type estimators of latent ability (Research Tech. Rep. No.82-30). Princeton, NJ: Educational Testing Service.

Karnjanawasri, S. (2002). *Modern test theories* (2 ed.). Bangkok, Thailand: Chulalongkorn University Press.

Karnjanawasri, S. (2005). *Classical test theory* (5th ed.). Bangkok, Thailand: Chulalongkorn University Press.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes towards computer-assted and other oral proficiency assessments. *Language Learning and Technology, 5*(2), 60-83.

Kyungsu, W. (1996). *Computerized adaptive testing : A comparison of item response theoretic approach and expert systems approach in polychotomous grading.* Unpublished PHD, Indiana University.

La-ongkaew, C. (1995). *A Construction of a computerized two stage test in mathematics for Prathom Suksa 5.* Unpublished M Ed, Khon Kaen University, Khon Kaen, Thailand.

Lawrence, R. (1998). *Item Banking. Practical Assessment, Research& Evaluation, 6(4).* Retrieved February, 14, 2007, from http://PAREonline.net/getvn.asp?v=6&n=4.

Leung Chi Keung, E. (2001). *Computerized adaptive testing as a means for mathematics assessment.* Retrieved 11 May, 2001, from http://www.fed.cuhk.edu.hk/~fllee/mathfor/edumath/9812/05leungck.html

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1-55.

Lila, S. (1996). *The developement of computerized item banking system.* Unpublished Doctor of Education, Srinakarinwirot University.

Linacre, J. M. (2005). The Partial Credit Model and the One-Item Rating Scale Model. *Rasch Measurement Transactions, 19*(1), 1001-1002.

Linden, W. J. v. d. (1999). Computerized educational testing. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment.* New York: Elsevier Science.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The developement and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 19*, 129-146.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, 7*, 1-84.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction testing and guidance.* New York: Harper and Row.

Lord, F. M. (1971a). Robins-monro procedures for tailored testing. *Journal of Educational and Psychological Measurement, 31*, 80-120.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

Lord, F. M. (1971c). A Theoretical study of two-stage testing. *Psychometricka, 36*, 277-242.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Apply Psychological Measurement, 1*, 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsale New Jersey: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test score.* Massachusetts: Addison-Wesley Publishing Company.

Lunz, M. E., & Bergstrom, B. A. (1991). Comparability of decissions for computer adaptive and written examinations. *Journal of Applied Health, 20*, 15-23.

Maneelek, R. (1997). *The effect of some variables on concurrent validity and item number of computerized adaptive testing.* Unpublished Ed D, Srinakharinwirot University, Bangkok, Thailand.

Master, G., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10*, 355-367.

Masters, G. N. (1997). Partial Credit Model. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement : an international handbook* (pp. 857-863). Cambridge, UK: Cambridge University Press.

Meejang, S., & Poonpun, S. (1999). *A meta - analysis of teaching and studying in mathematics for primary school in Thailand.* Bangkok: Department of Educational Research, Ministry of Education.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing : overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Mertens, D. M. (1998). *Research methods in education and psychology.* California: SAGE Publications, Inc.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of educational measurement, 21*(4), 315-330.

Ministry of Education. (2001). *Basic education curriculum.* Bangkok, Thailand: Ministry of Education.

Nakamura, Y. (2001). *Rasch measurement and item banking: Theory and practice* (Research Report No. 143). Washington, D.C.: ERIC Clearinghouse on Languages & Linguistics.

Nering, M. L. (1996). *The effect of person misfit in computerized adaptive testing.* Unpublished PhD, University of Minnesota.

Njiru, J. N. (2006). *Self-regulated learning in an ICT-Rich environment at a university.* Unpublished PhD thesis, The University of Western Australia.

Njiru, J. N., & Romanoski, J. (2007). *Development and calibration of Physics items to create an item bank, using a Rasch measurement model.* Paper presented at the The International Conference on Learning, Johannesburg, South Africa, June 26-29, 2007.

Office of the National Education Commission. (1999). *National education act of B.E.2542(1999).* Bangkok ,Thailand: Seven Printing Group.

Olsen, J. B., Maynes, D., Slawson, D., & Ho, K. (1986). *Comparison of paper-administered, computer-administered and computerized adaptive tests of achievement.* Paper presented at the the annual meeting of the American Educational Research Association, San francisco, CA.

Olsen, J. B., Maynes, D., Slawson, D., & Ho, K. (1986, April). *Comparison of paper-administered, computer-administered and computerized adaptive tests of achievement.* Paper presented at the the annual meeting of the American Educational Research Association, San francisco, CA.

Owen, R. J. (1969). *A Baysian approach to tailored testing (Research Bulletin No. 69-92).* Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Baysian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Paeratkool, C. (1975). *Measurement techniques [Thai Publication. Translation of Title]* (6 ed.). Bangkok: Wattanapanich.

Pallant, J. (2001). *SPSS survival manual: a step by step guide to data analysis using SPSS.* Crows Nest, NSW: Allen & Unwin.

Parshall, C. G. (2001). Automated test assembly for online administration. In C. G. Parshall, T. A. Davey, J. A. Spray & J. C. Kalohn (Eds.), *Practical conciderations in computer-based testing* (pp. 106-125). New York: Springer Verlag.

Patsula, L. N., & Steffen, M. (1997). *Maintaining item and test security in a cAT environment: A simulation study* (report No. 309). Chicago, IL: National Council on Measurement in Education.

Phungkham, N. (1988). *A comparison of the quality of CAT and Classicat Testing in English vocabulary ability of Mathayom Suksa 3 students.* Unpublished Master in Education, Chulalongkorn University, Thailand.

Pomsit, P. (2001). *An administration of the tailored test based on the Bayesian strategy in mathematics M. 014 by web page.* Unpublished M Ed, Khon Kaen University, Khon Kaen.

Punch, K. F. (1998). *Introduction on social research: Quatitative and quanlitative approaches*. London: Sage Publications.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denish Institute for Educational Research.

Rasch, G. (1980/1960). *Probabilistic model for some intelligence and attainment tests.* Chicago: The University of Chicago.

Richichi, R. V. (1996). *An analysis of test bank multiple-choice items using Item Response Theory* (Research /Technical No. ED 405367). USA.

Rudner, L. M. (1998a). *Item Banking* (No. EDO-TM-98-04). USA: Eric Clearinghouse on Assessment and Evaluation, Washington ,DC.

Rudner, L. M. (1998b). *Item banking. Practical Assessment, Researh& Evaluation, 6(4).* Retrieved June 9, 2007, from http://PAREonline.net/getvn.asp?v=6&n=4 .

Samejima, F. (1977). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38*, 221-233.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington DC: American Psychological Association.

Sangphueng, S., & Chooprateep, J. (2007). *Item bank in examination online system Chiangmai examination center.* Retrieved 20/05/2007, from http;//www.chiangmaiexam.com

Shermis, D., Stemmer, M., & Webb, M. (1996). Computerized adaptive skill assessment in a statewide testing program. *Journal of Research on Computing in Education, 29*, 49-67.

Skaggs, G., & Stevenson, J. (1989). A comparison of Pseudo-Bayesian and Joint Maximum Likelihood procedures for estimating item parameters in the three-parameter IRT model. *Applied Psychological Measurement, 13*(4), 391-402.

Songsaeng, K. (2004). *Test information functions in computerized adaptive testing.* Unpublished Ed.D Dissertation, Srinakharinwirot University, Bangkok, Thailand.

Songsang, K. (2004). *The Information Fuction of Computerized Adaptive Testing.* Unpublished Doctor of Education, Srinakharinwirot University.

Srisamran, P. (1997). *The evaluation of the project of item banking developement of the year1995-1997.* Sakonnakorn, Thailand: Sakonnakorn Area 1.

Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement, 22*, 271-279.

Styes, I., & Andrich, D. (1993). Linking the standard and advanced forms of Raven's Progressive Matrices in both the pencil-and-paper and cmputer-adaptive testing formats. *Educational and Psychological Measurement, 53*(4), 905-925.

Sukamolsan, S. (1996). *Item bank and computerized adaptive testing*. Bangkok ,Thailand: Wittayapat.

Sukamolson, S. (1996). *Item bank and computerized adaptive testing [Thai Publication. Translation of Title]*. Bangkok ,Thailand: Wittayapat.

Supeesut, N. (1998). *Construction of tailored test package and test administering through microcomputer in Mathematics for Mathayom Suksa 1 Level.* Unpublished Master Degree, ChiangMai University.

Supeesut, N. (1999). *Construction of tailored test package and test administering through microcomputer in Mathematics for Mathayom Suksa 1 Level.* Unpublished Master Degree, ChiengMai University.

Suwannoi, P. (1989). *A computerized peramidal testing in Chemistry for Mattayom Suksa 5 students.* Unpublished Master Degree, Khon Kaen University.

The Institute for the Promotion of Teaching Science and Technology. (2004, 24 May 2004). *News*. Retrieved August 26, 2004, from http://www.ipst.ac.th/news/May24news.html

Thissen, D., & Mislevy, R. J. (1990). Testing Algorithms. In H. Wainer (Ed.), *Computerized adaptive testing : A primer*. Hillsdale NJ: Erlbaum.

Tuntavanitch, P. (2006). *A Estimation of Elementary Students Efficacy in Division Skill using Computer Programing Method and Usual Testing Method.* Surin: Faculty of Education, Surin Rajabhat University, Thailand.

Umar, J. (1990). *Development of an examination system based on calibrated item bank networks* (Unpublished Project report): SIDEC, Standford University.

Umar, J. (1999). Item banking. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. New York: Pergamon Press.

van der Linden, W. J. (1999). Computerized educational testing. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 138-150). New York: PERGAMON.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of item response theory*. New York: Springer-Verlag.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Dordrecht, the Netherlands: Kluwer Academic.

Vicino, F. L., & Moreno, K. E. (1997). Human factors in the CAT system: a pilot study. In W. A. Sands, B. K. Waters & J. R.McBride (Eds.), *Computerized adaptive tesing: From inquiry to operation* (pp. 157-160). Washington, DC: American Psychological Association.

Wainer, H. (1990). Introduction and History. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg & D. Thissen (Eds.), *Computerized Adaptive Testing: A Primer* (pp. 1-22). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1993). Some practical considerations when covering a linearly administered test to an adaptive format. *Journal of Educational Measurement, 12*, 15-20.

Wainer, H., & Thissen, D. M. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*(4), 339-368.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45*, 373-391.

Waugh, R. F. (2002). Creating a scale to measure motivation to achieve academically: linking attitudes and behaviour using Rasch measurement. *British Journal of Educational Psychology, 72*(1), 65-86.

Waugh, R. F. (2003). *On the forefront of educational psychology*. New York: Nova Science Publishers.

Waugh, R. F. (2005). *Frontiers in educational psychology*. New York: Nova Science Publishers.

Waugh, R. F. (2006). Rasch measurement. In N. J. Salkind (Ed.), *The Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications.

Waugh, R. F., & Biswas, M. (2003). University acceptance of peers with disabilities. In R. F. Waugh (Ed.), *On the Forefront of Educational Psychology* (pp. 157-176). New York: Nova Science Publishers.

Waugh, R. F., Bowering, M. H., & Torok, S. (2005). Creating scales to measure reading comprehension and attitude and behaviour for prathom 7 students taught ESL through a genre-based method in Thailand. In R. F. Waugh (Ed.), *Frontiers in Educational Psychology* (pp. 133-174). (New York: Nova Science Publishers.

Waugh, R. F., Boyd, G., & Corrie, L. (2003). Teacher leadership in early childhood education: A Rasch measurement model analysis. In R. F. Waugh (Ed.), *On the*

Forefront of Educational Psychology (pp. 295-330). New York: Nova Science Publishers.

Waugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (True Score Theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement, 6*(1), 80-99.

Waugh, R. F., & Njiru, J. N. (2005). Measuring academic motivation to achieve for Malaysian high school students using a Rasch measurement model. In R. F. Waugh (Ed.), *Frontiers in Educational Psychology* (pp. 3-36). New York: Nova Science Publishers.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report No. 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 4*, 273-285.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70-84.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Acadamic Press.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement : Conventional or adaptive?* Minneapolis: University of Minnesota.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(Winter), 68-96.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of baysian adaptive testing. *Applied Psychological Measurement, 8*(3), 273-285.

Westers, P., & Kekderman, H. (1990). *Differentail item functioning in multiple choice items: Project psychometric aspects of item banking* (Research Report No. 47). Enschede, Netherlands: University of Twente.

Wiboonsri, Y. (2005). *Measurement and achievement test construction [Thai Publication. Translation of Title]* (4th ed.). Bangkok, Thailand: Chulalongkorn University Press.

Wise, S. L. (1997, March, 25-27). *Overview of practical issues in CAT program.* Paper presented at the The Annual Meeting of the National Council on Measurement in Education, Chicago.

Wolf, R. M. (1997). Questionnaires. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (2 ed.). New York: Pergamon Press.

Woodcock, R. W., & Johnson, M. B. (1977). *Woodcock-Johnson psycho-educational battery*. Chicago: Riverside.

Wright, B. D. (1999a). Fundamental measurement of psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. (1999b). Rasch measurement models. In G. N. Masters & J. P. Keeves (Eds.), *Advance in measurement in educational research and assessment* (pp. 85-97). Oxford, UK: Elsevier Science.

Wright, B. D., & Bell, S. R. (1984). Item banks: what, why, how. *Journal of Educational Measurement, 21*, 331-345.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago,IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, Illinois: Mesa Press.

APPENDICES

APPENDIX A

Questionnaire: Students Attitude towards Computerized Adaptive Testing

THIS QUESTIONNAIRE IS ANONYMOUS. PLEASE DON'T PUT YOUR NAME OR ANY IDENTIFICATION ON IT.

Direction: Please read all the following item wordings and answer by making a tick (√) in the box which best describes how strongly you agree or disagree with each wording. For example, if you strongly agree that computerized adaptive test is fair for all students, then √ the strongly agree box. Remember to √ one place for each item.

Item	Item Wording	Strongly disagree	Some-what disagree	Neither agree or	Some-what agree	Strongly agree
1	I am enthusiastic about taking part in a computerized adaptive test.					
2	I am happy and enjoyed doing a computerized adaptive test.					
3	The computerized adaptive testing is very interesting.					
4	I took the computerized adaptive test with confidence.					
5	I believe that I can do the computerized adaptive test well.					
6	I liked the computerized adaptive test because it was not too difficult for me.					
7	After finishing the computerized adaptive testing, I feel like wanting to do another.					
8	I like the computerized adaptive test because of its immediate feedback.					
9	I feel that it is worth having the chance to take the computerized adaptive test.					
10	I feel like I am using my full ability with the computerized adaptive test.					

Appendix A (continued)

Item Number	Item Wording	Strongly disagree	Some-what disagree	Neither agree or disagree	Some-what agree	Strongly agree
11	The computerized adaptive testing makes me want to study Mathematics.					
12	I feel lucky to have the chance to take a computerized adaptive test.					
13	I want a computerized adaptive testing to be used for other subjects.					
14	I am happy doing the computerized adaptive test without limited time.					
15	Computerized adaptive testing is very useful.					
16	Computerized adaptive testing is challenging.					
17	Computerized adaptive testing is modern.					
18	Computerized adaptive testing saves money.					
19	Computerized adaptive testing gives reliable results.					
20	The computerized adaptive testing is currently appropriate for these days.					
21	Computerized adaptive testing is fair for all students.					
22	Computerized adaptive testing inspires the students to do the test.					
23	Computerized adaptive testing allows students to spend less time on testing.					
24	Computerized adaptive testing provides examinees with appropriate items.					

Appendix A (continued)

Item Number	Item Wording	Strongly disagree	Some-what disagree	Neither agree or disagree	Some-what agree	Strongly agree
25	Computerized adaptive testing makes examinees careful when doing the test.					
26	I wish I could take a computerized adaptive test in a Mathematics test competition.					
27	I will tell my friends about computerized adaptive testing.					
28	If possible, I 'd rather take a computerized adaptive test.					
29	If I have a chance, I will introduce my younger friends to computerized adaptive testing.					
30	I am ready to apply the knowledge from computerized adaptive testing.					

APPENDIX B

Item difficulties from the Rasch analysis of the questionnaire data

Item Number	Item wording	Item difficulties
1	I am enthusiastic about taking part in a computerized adaptive test.	*+0.23*
2	I am happy and enjoyed doing a computerized adaptive test.	*+0.03*
3	The computerized adaptive testing is very interesting.	*-0.53*
4	I took the computerized adaptive test with confidence.	*+0.75*
5	I believe that I can do the computerized adaptive test well.	+0.72
6	I liked the computerized adaptive test because it was not too difficult for me.	+0.48
7	After finishing the computerized adaptive testing, I feel like wanting to do another.	+0.37
8	I like the computerized adaptive test because of its immediate feedback.	-0.11
9	I fell that it is worth having the chance to take the computerized adaptive test.	-0.06
10	I feel like I am using my full ability with the computerized adaptive test.	+0.01
11	The computerized adaptive testing makes me want to study Mathematics.	+0.28
12	I feel lucky to have the chance to take a computerized adaptive test.	+0.14
13	I want a computerized adaptive testing to be used for other subjects.	-0.41
14	I am happy doing the computerized adaptive test without limited time.	-0.29
15	Computerized adaptive testing is very useful.	-0.53
16	Computerized adaptive testing is challenging.	+0.05

Appendix B (continued)

Item Number.	Item wording	Item difficulties
17	Computerized adaptive testing is modern.	-0.73
18	Computerized adaptive testing saves money.	+0.61
19	Computerized adaptive testing gives reliable results.	-0.32
20	The computerized adaptive testing is currently appropriate for these days.	-0.56
21	Computerized adaptive testing is fair for all students.	-0.36
22	Computerized adaptive testing inspires the students to do the test.	+0.08
23	Computerized adaptive testing allows students to spend less time on testing.	+0.07
24	Computerized adaptive testing provides examinees with appropriate items.	+0.08
25	Computerized adaptive testing makes examinees careful when doing the test.	-0.00
26	I wish I could take a computerized adaptive test in a Mathematics test competition.	+0.33
27	I will tell my friends about computerized adaptive testing.	+0.36
28	If possible, I 'd rather take a computerized adaptive test.	-0.13
29	If I have a chance, I will introduce my younger friends to computerized adaptive testing.	-0.11
30	I am ready to apply the knowledge from computerized adaptive testing.	-0.43

APPENDIX C

Item locations (78 items), SE, Residuals, fit to the model and Probability order of mathematics test

Item Number	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob
76	+0.21	0.12	+1.55	364.84	396	3.46	0.94
60	+0.95	0.12	+0.94	366.68	398	3.62	0.93
73	+0.85	0.12	+0.85	361.15	392	4.41	0.88
87	-0.66	0.12	-0.25	364.84	396	4.54	0.87
103	-0.12	0.12	-0.04	354.70	385	4.59	0.87
127	-0.07	0.12	+2.26	349.18	379	4.84	0.84
59	-1.27	0.12	-0.38	366.68	398	4.85	0.84
233	+1.37	0.13	-0.39	323.38	351	4.85	0.84
150	-0.28	0.12	+0.76	329.83	358	5.01	0.83
202	-0.45	0.13	-0.35	304.95	331	5.09	0.82
97	+1.37	0.13	+0.24	354.70	385	5.11	0.82
45	-0.45	0.13	+0.34	320.62	348	4.77	0.78
200	-0.70	0.13	+0.42	303.11	329	4.91	0.76
220	-0.50	0.13	+0.13	322.46	350	5.77	0.76
81	+0.28	0.12	+2.24	365.76	397	6.08	0.72
249	-0.09	0.12	+0.97	322.46	350	6.08	0.72
92	-0.33	0.12	+1.64	355.62	386	6.21	0.71
29	+0.14	0.12	+0.72	319.69	347	5.39	0.71
100	-0.02	0.12	-0.68	356.55	387	6.34	0.70
138	-1.12	0.14	-0.62	330.75	359	6.48	0.68
75	+0.63	0.12	+1.07	363.92	395	6.58	0.67
171	+0.12	0.13	-0.33	302.19	328	5.73	0.67
129	+0.10	0.12	+2.35	351.94	382	6.72	0.67
106	+0.24	0.12	+0.27	355.62	386	7.15	0.61
51	-0.85	0.12	+0.40	364.84	396	7.19	0.61
248	-0.03	0.12	+0.50	323.38	351	7.21	0.60
126	-0.07	0.12	-0.94	355.62	386	7.25	0.60
174	-0.83	0.14	-0.28	302.19	328	7.25	0.60
214	-1.07	0.14	-1.17	320.62	348	7.27	0.60
53	-0.62	0.12	+1.65	364.84	396	7.73	0.55
164	+1.01	0.13	+1.80	326.14	354	7.93	0.53
122	-0.62	0.12	+1.13	355.62	386	8.47	0.47
68	-0.04	0.12	+1.05	363.00	394	8.67	0.45
13	-0.05	0.12	+1.19	320.62	348	7.72	0.45
116	-0.77	0.12	+1.36	356.55	387	8.94	0.43
104	+0.24	0.12	-0.92	353.78	384	9.00	0.42
125	-0.21	0.12	+0.39	355.62	386	9.12	0.41
85	-0.32	0.11	+1.87	364.84	396	9.15	0.41
90	+0.16	0.12	-0.10	365.76	397	9.41	0.38
112	+1.08	0.13	-0.23	355.62	386	9.47	0.38
48	+0.22	0.12	-0.24	319.69	347	8.52	0.37

Appendix C (continued)

Item Number	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob
130	-0.05	0.12	+0.65	354.70	385	9.65	0.36
156	+0.67	0.12	+1.35	329.83	358	9.72	0.36
197	-0.06	0.13	+2.71	303.11	329	9.72	0.36
182	+0.73	0.13	-1.01	304.03	330	9.80	0.35
34	+1.57	0.14	-0.08	317.85	345	8.87	0.34
71	+0.68	0.12	-0.21	364.84	396	10.11	0.32
132	-0.10	0.12	+2.28	326.14	354	10.11	0.32
139	-0.93	0.13	-0.47	329.83	358	10.19	0.32
157	-0.41	0.13	+1.75	327.06	355	9.41	0.29
218	-0.48	0.13	+0.94	323.38	351	10.59	0.28
14	-0.66	0.13	-0.41	318.77	346	8.34	0.28
91	-0.61	0.12	+0.28	356.55	387	10.61	0.28
74	+0.22	0.12	-1.00	363.00	394	10.77	0.27
223	+0.29	0.12	+0.16	322.46	350	11.09	0.25
119	-0.82	0.12	-1.20	353.78	384	11.51	0.22
142	+0.37	0.12	-1.61	330.75	359	11.64	0.21
183	+0.54	0.13	+0.57	304.03	330	11.67	0.21
224	+0.19	0.12	-0.48	323.38	351	11.80	0.20
141	-0.13	0.12	-0.86	329.83	358	11.94	0.19
135	-0.71	0.13	+0.49	328.91	357	12.65	0.15
185	+0.27	0.13	+0.34	304.95	331	12.91	0.14
140	+0.47	0.12	-0.56	330.75	359	13.07	0.13
172	+0.39	0.13	+0.89	304.03	330	13.06	0.13
134	+0.85	0.13	+2.53	327.06	355	13.37	0.12
212	+0.27	0.12	+2.20	323.38	351	13.46	0.12
117	+0.25	0.12	+2.22	356.55	387	14.36	0.08
47	+0.14	0.12	+1.51	319.69	347	13.33	0.08
109	-0.15	0.12	-0.29	356.55	387	14.63	0.07
115	+0.08	0.12	+3.19	356.55	387	14.93	0.07
170	+0.14	0.12	+2.03	328.91	357	15.74	0.04
121	+0.65	0.12	+2.45	354.70	385	16.89	0.02
241	-0.54	0.13	+1.10	323.38	351	17.57	0.01
83	+0.74	0.12	+2.20	363.92	395	18.43	0.00
16	-0.96	0.13	-1.03	320.62	348	21.24	0.00
88	-0.86	0.12	-1.79	364.84	396	20.37	0.00
180	-0.01	0.13	+4.01	304.03	330	29.45	0.00
236	+0.51	0.12	+2.90	321.54	349	18.53	0.00

APPENDIX D

Item locations (20 items), SE, Residuals, fit to the model and Probability order of mathematics test

Item Number	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob
40	+0.18	0.10	+1.04	559.33	590	1.81	0.99
42	-0.15	0.09	+0.90	568.81	600	3.10	0.93
46	+0.42	0.10	-0.41	565.02	596	3.51	0.90
22	+0.51	0.10	+1.15	569.76	601	4.88	0.76
45	+0.12	0.09	+0.89	565.02	596	6.68	0.56
36	+0.15	0.09	+1.86	568.81	600	7.10	0.51
17	. -0.46	0.09	+1.86	568.81	600	7.45	0.48
12	-0.29	0.09	+1.73	567.87	599	7.69	0.45
37	+0.09	0.09	+2.19	568.81	600	8.91	0.33
41	+0.12	0.09	+0.34	568.81	600	8.90	0.33
20	-0.11	0.09	-0.25	564.07	595	9.18	0.31
7	-1.65	0.09	+0.25	570.71	602	10.10	0.24
31	+0.43	0.10	+0.12	570.71	602	10.22	0.23
23	+0.39	0.10	+1.68	561.23	592	10.84	0.19
24	+0.11	0.09	-0.25	570.71	602	10.86	0.19
47	-0.23	0.09	-0.44	567.87	599	11.53	0.15
50	+0.31	0.10	+1.51	569.76	601	11.72	0.14
18	-0.53	0.09	+0.06	565.02	596	12.05	0.13
32	+0.28	0.10	+0.54	567.87	599	14.44	0.05
43	+0.31	0.10	-0.62	565.97	597	17.38	0.00

APPENDIX E

Item thresholds (30 items) of the questionnaire of student attitude towards computerized adaptive testing

Item Number	Mean	THRESHOLDS		
		1	2	3
1	+0.23	-1.30	+0.44	+1.56
2	+0.03	-1.14	-0.09	+1.30
3	-0.53	-1.28	-0.76	+0.45
4	+0.75	-0.78	+0.98	+2.04
5	+0.72	-0.93	+0.90	+2.18
6	+0.48	-0.71	+0.76	+1.40
7	+0.37	-0.36	+0.19	+1.28
8	-0.11	-0.84	-0.05	+0.56
9	-0.06	-0.90	-0.31	+1.04
10	+0.01	-1.35	+0.03	+1.35
11	+0.28	-1.15	+0.43	+1.54
12	+0.14	-0.81	-0.09	+1.32
13	-0.41	-1.06	-0.37	+0.20
14	-0.29	-0.86	-0.04	+0.02
15	-0.53	-1.85	-0.38	+0.65
16	+0.05	-1.34	+0.01	+1.48
17	-0.73	-1.40	-1.04	+0.24
18	+0.61	-0.60	+0.68	+1.73
19	-0.32	-1.90	-0.06	+0.98
20	-0.56	-1.52	-0.60	+0.43
21	-0.36	-1.69	-0.24	+0.86
22	+0.08	-1.06	-0.13	+1.44
23	+0.07	-1.15	+0.29	+1.07
24	+0.08	-0.95	-0.20	+1.38
25	-0.00	-1.71	+0.32	+1.38
26	+0.33	-0.99	+0.67	+1.29
27	+0.36	-0.87	+0.49	+1.46
28	-0.13	-1.09	-0.08	+0.80
29	-0.11	-1.39	-0.03	+1.10
30	-0.43	-2.04	-0.19	+0.95

APPENDIX F

Item locations (30 items), SE, Residuals and fit to the model of the questionnaires of students attitude towards computerized adaptive testing

Item Number	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob
1	+0.23	0.07	+1.91	382.66	399	5.59	0.33
2	+0.03	0.07	-0.15	382.66	399	6.72	0.22
3	-0.53	0.08	-0.88	380.74	397	5.58	0.33
4	+0.75	0.07	+2.49	379.78	396	8.69	0.10
5	+0.72	0.07	+0.60	381.70	398	1.78	0.88
6	+0.48	0.06	-0.51	380.74	397	2.93	0.70
7	+0.37	0.06	+2.18	378.82	395	8.67	0.10
8	-0.11	0.07	+1.37	380.74	397	13.15	0.00
9	-0.06	0.07	-2.63	381.70	398	16.23	0.00
10	+0.01	0.07	+0.30	381.70	398	1.07	0.96
11	+0.28	0.07	-1.04	381.70	398	3.03	0.69
12	+0.14	0.07	-0.60	381.70	398	3.42	0.62
13	-0.41	0.07	-0.73	381.70	398	0.46	0.99
14	-0.29	0.07	+1.14	382.66	399	6.66	0.22
15	-0.53	0.07	+0.05	380.74	397	3.06	0.68
16	+0.05	0.07	+1.87	382.66	399	9.64	0.06
17	-0.73	0.08	-0.61	382.66	399	4.23	0.50
18	+0.61	0.06	+3.16	380.74	397	10.14	0.04
19	-0.32	0.07	+0.28	380.74	397	0.42	0.99
20	-0.56	0.08	-1.58	381.70	398	4.32	0.49
21	-0.36	0.07	-0.19	381.70	398	2.45	0.78
22	+0.08	0.07	+0.34	380.74	397	1.43	0.92
23	+0.07	0.07	+1.72	377.86	394	12.23	0.00
24	+0.08	0.07	-0.94	380.74	397	6.64	0.23
25	-0.00	0.07	-0.36	376.91	393	3.40	0.63
26	+0.33	0.06	-0.82	381.70	398	3.00	0.69
27	+0.36	0.06	+0.56	380.74	397	7.56	0.16
28	-0.13	0.07	-1.29	374.99	391	7.78	0.14
29	-0.11	0 07	-0.40	381.70	398	2.44	0.78
30	-0.43	0.07	-0.07	382.66	399	2.72	0.74

Note

1. Location means the item difficulty.
2. SE means the standard error.
3. Residual means the item-person interaction test of fit statistic for each item.
4. DegFree means the degree of freedom associated with the Residual value.
5. DatPts means the number of persons associated with an item.
6. Chi Sq means the item-trait interaction chi square statistics.
7. Prob means probability.
8. All numbers are given to two decimal places because the errors are between 0.06 and 0.08 logits.

APPENDIX G

Item locations, SE, Residuals, fit to the model and Probability order of the questionnaires of students attitude towards computerized adaptive testing

Item Number	Location	SE	Residual	DegFree	DatPts	Chi Sq	Prob
19	-0.32	0.07	+0.28	380.74	397	0.42	0.99
13	-0.41	0.07	-0.73	381.70	398	0.46	0.99
10	+0.01	0.07	+0.30	381.70	398	1.07	0.96
22	+0.08	0.07	+0.34	380.74	397	1.43	0.92
5	+0.72	0.07	+0.60	381.70	398	1.78	0.88
29	-0.11	0 07	-0.40	381.70	398	2.44	0.78
21	-0.36	0.07	-0.19	381.70	398	2.45	0.78
30	-0.43	0.07	-0.07	382.66	399	2.72	0.74
6	+0.48	0.06	-0.51	380.74	397	2.93	0.70
26	+0.33	0.06	-0.82	381.70	398	3.00	0.69
11	+0.28	0.07	-1.04	381.70	398	3.03	0.69
15	-0.53	0.07	+0.05	380.74	397	3.06	0.68
25	-0.00	0.07	-0.36	376.91	393	3.40	0.63
12	+0.14	0.07	-0.60	381.70	398	3.42	0.62
17	-0.73	0.08	-0.61	382.66	399	4.23	0.50
20	-0.56	0.08	-1.58	381.70	398	4.32	0.49
3	-0.53	0.08	-0.88	380.74	397	5.58	0.33
1	+0.23	0.07	+1.91	382.66	399	5.59	0.33
24	+0.08	0.07	-0.94	380.74	397	6.64	0.23
14	-0.29	0.07	+1.14	382.66	399	6.66	0.22
2	+0.03	0.07	-0.15	382.66	399	6.72	0.22
27	+0.36	0.06	+0.56	380.74	397	7.56	0.16
28	-0.13	0.07	-1.29	374.99	391	7.78	0.14
7	+0.37	0.06	+2.18	378.82	395	8.67	0.10
4	+0.75	0.07	+2.49	379.78	396	8.69	0.10
16	+0.05	0.07	+1.87	382.66	399	9.64	0.06
18	+0.61	0.06	+3.16	380.74	397	10.14	0.04
23	+0.07	0.07	+1.72	377.86	394	12.23	0.00
8	-0.11	0.07	+1.37	380.74	397	13.15	0.00
9	-0.06	0.07	-2.63	381.70	398	16.23	0.00

APPENDIX H

Letters Seeking Permission and Consent Form

A letter to Student.

Ubon Ratchathani Rajabhat
University, Muang District,
Ubon Ratchathani, 34000,
Thailand.
Telephone (045) 352000-29
Facsimile (045) 311465

10 March 2004

Dear Student

I wish to request your participation in a study I am conducting focusing on computerized adaptive testing in Mathematics for Prathom Suksa 6 students. Please express your agreement to participate by signing the consent form below. The test is not a part of your study. The score you gain from taking the test cannot be useful for any subject you are studying. You have the rights to quit at anytime you would like. The information you provide will be very useful in constructing an innovation in learning assessment and evaluation. Please indicate your willingness to participate.

Thank you for your participation and co-operation.

Sincerely yours,

(Assistant Professor Chaowprapha Chuesathuchon)

Faculty of Education, Ubon Ratchathani Rajabhat University
Muang District, Ubon ratchathani, 34000, Thailand.

Please put a tick in one of the boxes below

I have learned about the details of this study and my rights to quit whenever I want to.

☐ I wish to participate in this study. ☐ I do not wish to participate in this study.

I also understand that my identity will remain anonymous; and that my grade will not depend on whether or not I take part in the study.

Signature (student) _____

Signature (parents / guardian) _____

Date_____ /_____/2004

A Letter of Seeking Permission

Edith Cowan University
1 February 2004

Dear School Director,

Subject: Seeking permission to conduct a research project

Further to my university approved research project entitled "Computerized Adaptive Testing in Mathematics for Primary Schools in Thailand", I would like to ask for your permission to carry out research in your school. This study aims to achieve understanding about Prathom Suksa 6 student abilities in mathematics and attitudes to the Mathematics Computerized Adaptive Testing. This information will lead to develop a new innovation in learning assessment and evaluation in Thailand. The Prathom Suksa 6 students enrolling in the academic year 2004, have been selected to be subject of this study.

Your approval and support would be highly appreciated.

Sincerely yours,

(Assistant Professor Chaowprapha Chuesathuchon)

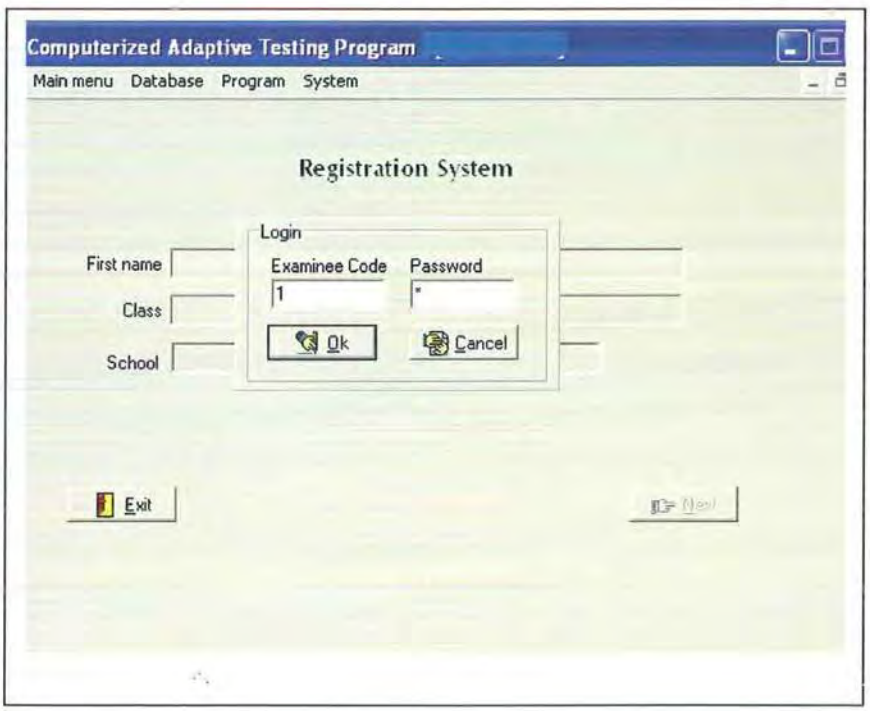
Enclosures (2): 1. Ethics clearance
2. Research proposal

APPENDIX I

The process of Computerized Adaptive Testing.

1. An examinee types an examinee code and password in blanks of registration form.

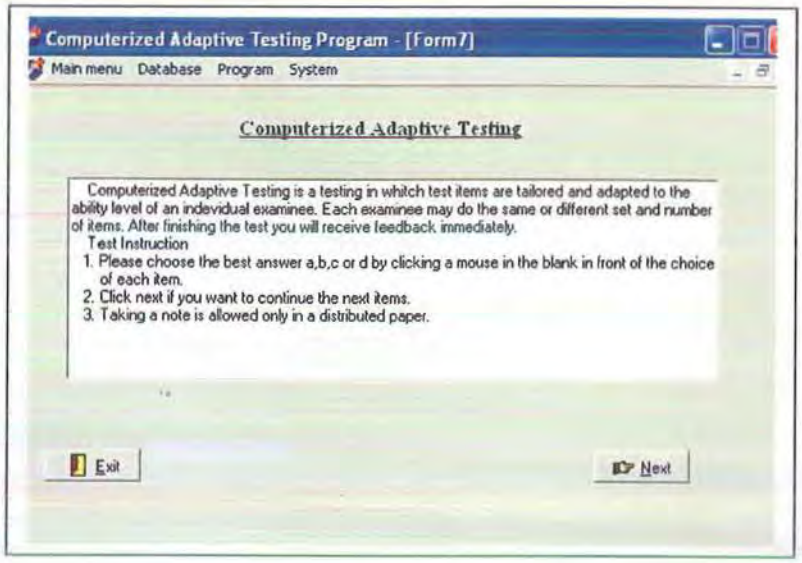
2. After the examinee completed typing code and password he/she then clicked ok.



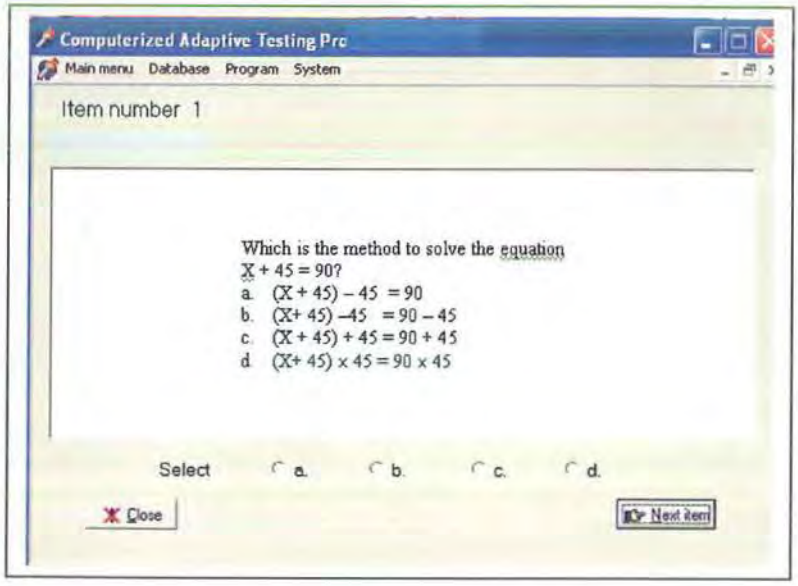
3. After clicking ok, the examinee's information was shown on the screen.

4. The examinee then clicked yes to confirm the information.

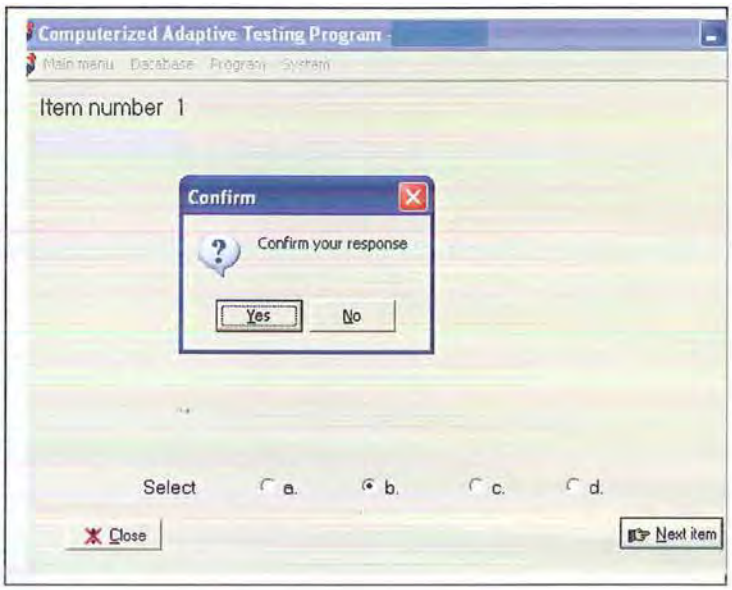
5. After that the test instruction was shown on the screen, the examinee then read the test instruction and clicked "next' button to start the test.



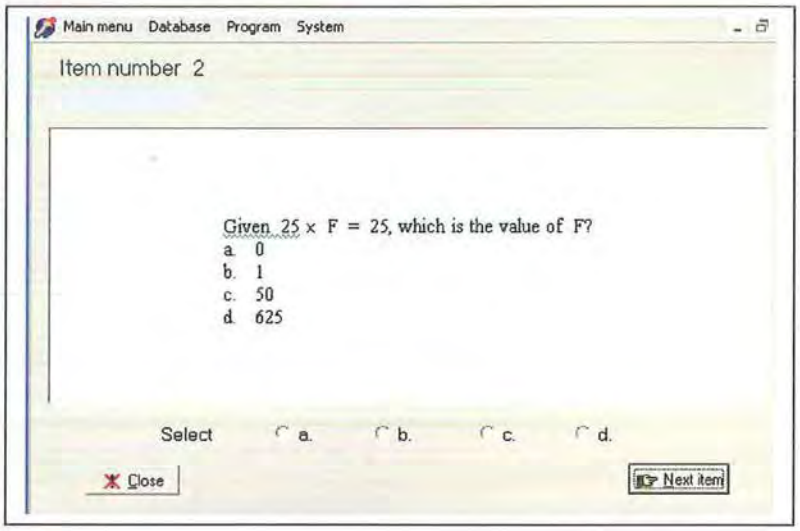
6. The initial item (item number 1) was randomised as a meduim difficulty item and displayed.



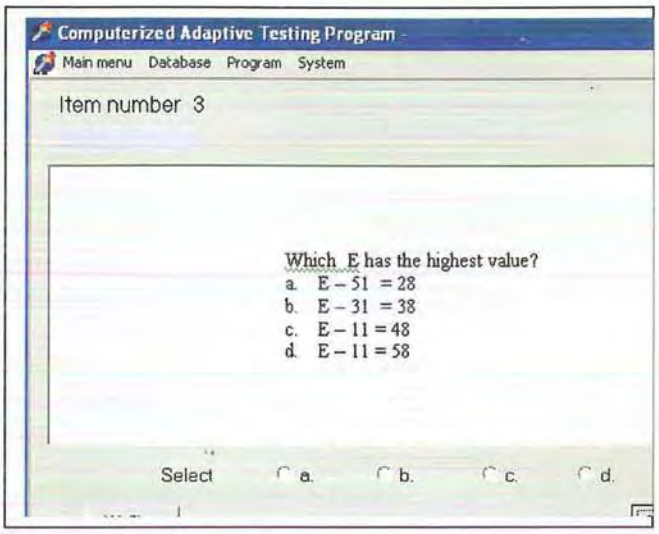
7. The examinee chose the best answer a, b, c, or d, by clicking a mouse in the blank in front of the choice at then clicked yes to confirm the response .

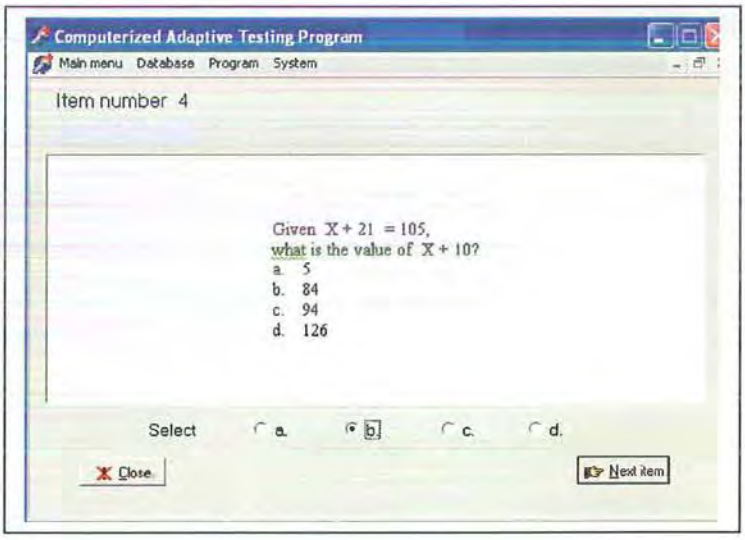


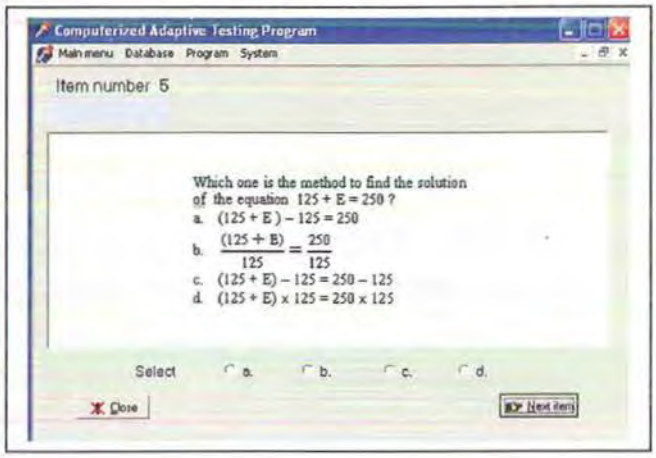
8. After that item number 2 was displayed, the examinee chose the best answer a, b, c, or d, by clicking a mouse in the blank in front of the choice at then clicked yes to confirm the response as same as in item 1.

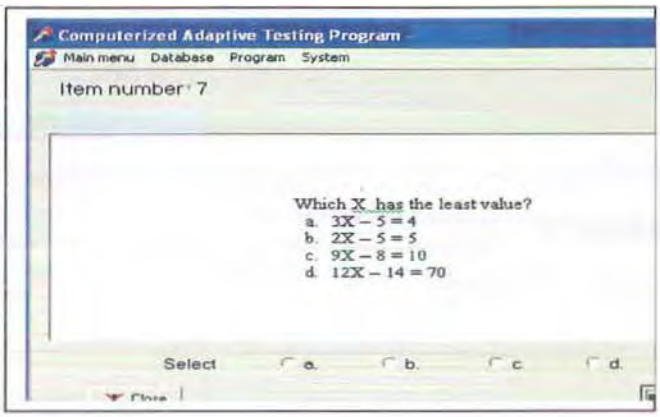


9. The processing was continued until the last item was answered (item 7 for this case) .









10. After the completion of the test, the results of the examinee (true score and ability estimate) was automatically recorded in computer and also shown on the screen.