COMPARISON OF LINEAR AND ADAPTIVE VERSIONS OF THE TURKISH PUPIL MONITORING SYSTEM (PMS) MATHEMATICS ASSESSMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEMİRHAN GÖKÇE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
DEGREE OF DOCTOR OF PHILOSOPHY
IN SECONDARY SCIENCE AND MATHEMATICS EDUCATION

JULY 2012

Approval of the thesis:

COMPARISON OF LINEAR AND ADAPTIVE VERSIONS OF THE TURKISH PUPIL MONITORING SYSTEM (PMS) MATHEMATICS ASSESSMENT

submitted by **SEMİRHAN GÖKÇE** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Secondary Science and Mathematics Education Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Ömer Geban _____
Head of Department, **Secondary Science and Mathematics Education**

Prof. Dr. Giray Berberoğlu _____
Supervisor, **Secondary Science and Mathematics Educ. Dept., METU**

Examining Committee Members:

Prof. Dr. Ömer Geban _____
Secondary Science and Mathematics Educ. Dept., METU

Prof. Dr. Giray Berberoğlu _____
Secondary Science and Mathematics Educ. Dept., METU

Prof. Dr. Nükhet Demirtaşlı _____
Educational Sciences Dept., Ankara University

Assoc. Prof. Dr. Esen Uzuntiryaki _____
Secondary Science and Mathematics Educ. Dept., METU

Assoc. Prof. Dr. Yezdan Boz _____
Secondary Science and Mathematics Educ. Dept., METU

Date: *20.07.2012*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Semirhan Gökçe

Signature : _____

ABSTRACT

COMPARISON OF LINEAR AND ADAPTIVE VERSIONS OF THE TURKISH PUPIL MONITORING SYSTEM (PMS) MATHEMATICS ASSESSMENT

Gökçe, Semirhan

Ph. D., Department of Secondary Science and Mathematics Education

Supervisor: Prof. Dr. Giray Berberoğlu

July 2012, 131 pages

Until the developments in computer technology, linear test administrations within classical test theory framework is mostly used in testing practices. These tests contain a set of predefined items in a large range of difficulty values for collecting information from students at various ability levels. However, placing very easy and very difficult items in the same test not only cause wasting time and effort but also introduces possible extraneous variables into the measurement process such as possibility of guessing, chance of careless errors induced by boredom or frustration. Instead of administering a linear test there is another option that adapts the difficulty of test according to the ability level of examinees which is named as computerized adaptive test. Computerized adaptive tests use item response theory as a measurement framework and have algorithms responsible for item selection, ability estimation, starting rule and test termination.

The present study aims to determine the applicability of computerized adaptive testing (CAT) to Turkish Pupil Monitoring System's (PMS) mathematics assessments. Therefore, live CAT study using only multiple choice items is designed to investigate whether to obtain comparable ability estimations. Afterwards, a Monte Carlo simulation study and a Post-hoc simulation study are designed to determine the optimum CAT algorithm for Turkish PMS mathematics assessments. In the simulation studies, both multiple-choice and open-ended items are used and different scenarios are tested regarding various starting rules, termination criterion, ability estimation methods and existence of exposure/content controls.

The results of the study indicate that using Weighted Maximum Likelihood (WML) ability estimation method, easy initial item difficulty as starting rule and a fixed test reliability termination criterion (0.30 standard error as termination rule) gives the optimum CAT algorithm for Turkish PMS mathematics assessment. Additionally, item exposure and content control strategies have a positive impact on providing comparable ability estimations.

Keywords: *Pupil monitoring system, computerized adaptive testing, ability estimation method, termination criteria, item exposure control and content control strategies*

ÖZ

TÜRKİYE ÖĞRENCİ İZLEME SİSTEMİ (ÖİS) MATEMATİK DEĞERLENDİRMESİNİN BİLGİSAYAR ORTAMINDAKİ LİNEER VE BİREYSELLEŞTİRİLMİŞ TEST VERSİYONLARININ KARŞILAŞTIRILMASI

Gökçe, Semirhan

Doktora, Ortaöğretim Fen ve Matematik Alanları Eğitimi Bölümü

Tez Yöneticisi: Prof. Dr. Giray Berberoğlu

Temmuz 2012, 131 sayfa

Klasik test kuramı çatısını kullanan lineer testler bilgisayar teknolojisindeki gelişmelerden önce test uygulamalarında sıklıkla kullanılmaktaydı. Uygulama öncesinde katılımcılara hangi soruların yöneltileceğinin bilindiği bu testlerde öğrencilerin yetenek düzeyindeki çeşitlilik nedeniyle farklı zorluk düzeyinde birçok soru bulunmaktadır. Klasik kağıt-kalem uygulamasında, öğrencinin düzeyine uygun olmayan çok kolay ve çok zor soruların aynı testte kullanılması zaman ve işgücü kaybını da beraberinde getirmekte öğrencilerin yanıtı tahminle bulmasına olanak sağladığı gibi sınav esnasında öğrencinin konsantrasyon kaybına da yol açabilmektedir. Lineer test uygulamasından farklı olarak öğrencinin yetenek düzeyine uygun güçlükte soru yöneltilmesi bilgisayar ortamında bireyselleştirilmiş (BOB) testlerin temel prensibi olarak tanımlanabilir.

Madde Tepki Kuramı (MTK)'nı kullanan bu testlerde yetenek düzeyinin tahmin edilmesi; madde seçimi, yetenek kestirim yöntemi, test başlangıç ve sonlandırma kriterlerinin tanımlandığı algoritmalar ile mümkündür.

Bu çalışmanın amacı Öğrenci İzleme Sistemi (ÖİS) matematik değerlendirmesinin bilgisayar ortamında bireyselleştirilmiş test olarak uygulanabilirliğini araştırmaktır. Bu nedenle çoktan seçmeli soruların kullanıldığı gerçek BOB test uygulaması tasarlanmış ve ÖİS sonuçlarıyla karşılaştırılabilir sonuçların alınıp alınmadığı incelenmiştir. Sonrasında ise çoktan seçmeli ve açık uçlu soru verilerinin kullanıldığı Monte Carlo ve Post-hoc simülasyonları kullanılarak birçok senaryo tasarlanmış ve en uygun BOB test uygulama algoritması belirlenmiştir. Bu simülasyonlar test başlangıç ve sonlandırma kriterleri, yetenek kestirim yöntemleri ve açığa çıkma/içerik kontrol stratejileri olarak belirtilebilir.

Çalışmadan elde edilen bulgulara gore ÖİS matematik değerlendirmesi için en uygun algoritmanın Weighted Maximum Likelihood (WML) yetenek kestirim yönetimi olduğu, testin kolay soruyla başlandığı, yetenek kestiriminin 0.30 standart hata ile hesaplanınca testin sonlandığı ve açığa çıkma/içerik kontrol stratejilerinin her ikisinin birden kullanılması olduğu belirlenmiştir.

Anahtar Kelimeler: *Öğrenci izleme sistemi, bilgisayar ortamında bireyselleştirilmiş test, yetenek kestirim yöntemi, test sonlandırma kriteri, açığa çıkma ve içerik kontrol stratejileri*

To my family

ACKNOWLEDGMENTS

First of all, the author is highly indebted to his supervisor Prof. Dr. Giray Berberoğlu for his guidance, advices, criticism and encouragements throughout the study. Indeed, his detailed comments and insight have been a great value for the author in finding solutions to the problems he confronts.

Moreover, Cito Türkiye has an important role in this study. The author gives his special thanks to Ömer Konak and Çiğdem İş Güzel for their trust and support on using Pupil Monitoring Sytem (PMS) data. The technical assistance of Derya Kaymak during the implementation of computerized adaptive testing (CAT) is gratefully acknowledged.

The author really appreciates the contributions of İlker Kalender because of the truly sharing his experiences about CAT algorithm during the study.

Furthermore, it is the author's chance to meet with Prof. Dr. Cees Glas and Prof. Dr. Theo Eggen during his internship at Research Methodology, Measurement and Data Analysis Department of University of Twente in the Netherlands. The author greatly expresses his gratitude especially to Mr. Glas for his support and guidance especially in the Item Response Theory (IRT) framework, data analysis, implementation of CAT and the simulation studies. The author also would like to thank Mr. Eggen because he provides extra information about the related studies and shares his ideas about this study.

Finally, the author also expresses his love and deepest feelings to his beloved family and Tülay Acaray for their support and motivation.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ vi
ACKNOWLEDGMENTS ix
TABLE OF CONTENTS x
LIST OF TABLES xii
LIST OF FIGURES xvi
LIST OF ABBREVIATIONS xvii

CHAPTERS
1. INTRODUCTION 1
1.1 Computerized Adaptive Testing: Past and Present 3
1.2 Advantages and Limitations of Computerized Adaptive Testing 5
1.3 Pupil Monitoring System (PMS) 8
1.4 Turkish Pupil Monitoring System 9
1.5 CAT Administration in Pupil Monitoring System 10
1.6 Definition of Terms 11
1.7 Purpose of the Study 12
1.8 Significance of the Study 14

2. LITERATURE REVIEW 16
2.1 Basics of Item Response Theory (IRT) 16
2.2 Studies Related to IRT 28
2.3 Basics of Computerized Adaptive Testing (CAT) 31
2.4 Studies Related to CAT 38
2.5 Summary of the Literature Review 48

3. METHODOLOGY 50
3.1 Sample of the Study 51
3.2 Model Data Fit 55
3.3 Live CAT Administration 60
3.4 Monte Carlo Simulations 61
3.5 Post-hoc Simulations 61

4. RESULTS 63
4.1 Results of Live CAT Study 63
4.2 Results of Monte Carlo Simulation Studies 65
4.3 Results of Post-hoc Simulation Studies 73

5. CONCLUSION AND DISCUSSION 85
5.1 Summary of Findings 85
5.2 Live CAT Administration Phase 87
5.3 Monte Carlo Simulations Phase 88
5.4 Post-hoc Simulations Phase 90
5.5 Applicability of CAT in Turkish PMS 92
5.6 Limitations of the Study 93
5.7 Suggestions for Further Research 94

REFERENCES 96

APPENDIX A: ITEM PARAMETERS CALIBRATED BY BILOG-MG, OPLM AND MIRT 106

APPENDIX B: CORRELATION COEFFICIENTS OF ABILITY ESTIMATIONS WITH TURKISH PMS SUB-DOMAIN SCORES IN DIFFERENT SCENARIOS 127

CURRICULUM VITAE 131

LIST OF TABLES

TABLES

Table 3.1 Distribution of Items used in Live CAT Administration Regarding Content Area …… 51

Table 3.2 Summary Statistics for PMS Item Parameters Calibrated by BILOG-MG …… 52

Table 3.3 Number of Items Used in Simulation Phase Regarding Content Area and Item Types……. 53

Table 3.4 Summary Statistics for PMS Item Parameters Calibrated by MIRT …….. 54

Table 3.5 Correlation Coefficients of Item Discrimination Parameters Obtained from OPLM, BILOG-MG and MIRT ……. 55

Table 3.6 Correlation Coefficients of Item Difficulty Parameters Obtained from OPLM, BILOG-MG and MIRT ……. 55

Table 3.7 Eigenvalues of Factor Analysis Results …… 56

Table 3.8 Percentages of Low Ability Students' Correct Responses on 10 Most Difficult Items ……. 57

Table 3.9 Percentages of Missing Responses on Last 5 Items in Tasksets…… 59

Table 4.1 Correlation Coefficients of Ability Estimations between PMS Computer Based Linear Mathematics Assessments and Live CAT Administration ……..…… 64

Table 4.2 Percentages of Correct Proficiency Levels in Live CAT Study ……..... 64

Table 4.3 Number of Items Administered under Different Starting Rules and Fixed Test Reliability Termination Rules Regarding WML and EAP Ability Estimations……. 66

Table 4.4 Correlation Coefficients between Expected and Observed Ability Estimations of Different Starting Rules and Fixed Test ReliabilityValues in WML and EAP Ability Estimation Methods …… 67

Table 4.5 Standard Error Estimations of Monte Carlo Simulations under Different Starting Rules and Fixed Test Length Termination Rules by WML and EAP Ability Estimation Methods……69

Table 4.6 Correlation Coefficients between Expected and Observed Scores of Different Starting Rules and Fixed Test Length Values by WML and EAP Ability Estimation Methods……. 70

Table 4.7 Results of 10 Replications in WML to Determine Item Exposure Rates…… 72

Table 4.8 Results of 10 Replications in EAP to Determine Item Exposure Rates……..... 72

Table 4.9 Number of Items Administered under Different Starting Rules and Fixed Test Reliability Termination Rules in Maximum Likelihood (ML) Ability Estimation…… 74

Table 4.10 Correlations Coefficients of PMS Mathematics Assessment Scores under Different Starting Rules and Fixed Test Reliability Termination Rules…… 75

Table 4.11 Standard Error Estimations of Post-hoc Simulations under Different Starting Rules and Fixed Test Length Termination Rules in ML Ability Estimation……. 77

Table 4.12 Correlations Coefficients of PMS Mathematics Assessment Scores under Different Starting Rules and Fixed Test Length Termination Rules...…….. 78

Table 4.13 Content Analysis of Items in CAT Simulation ……..….. 80

Table 4.14 Correlations of Ability Estimations between PMS Assessment and Post-Hoc CAT Simulation under the Use of Content and Exposure Control…… 81

Table 4.15 Correlation Coefficients of Ability Estimations between PMS Mathematics Assessments and Optimum CAT Algorithm Simulation Results……...… 83

Table 4.16 Percentages of Correct Proficiency Levels with Different Termination Criteria in Post-hoc Simulations …….. 84

Table A.1 Item Parameters Calibrated by BILOG-MG, OPLM and MIRT …… 107

Table B.1 Correlation Coefficients of PMS Mathematics Assessments Sub-Domain Scores with ML and WML Ability Estimations of Post-hoc CAT Simulations under Different Starting and Fixed Test Reliability Termination Rules…..…….. 128

Table B.2 Correlation Coefficients of PMS Mathematics Assessments Sub-Domain Scores with ML and WML Ability Estimations of Post-hoc CAT Simulations under Different Starting and Fixed Test Length Termination Rules…… 129

Table B.3 Correlation Coefficients of Ability Estimations between PMS Mathematics Assessment and Post-Hoc CAT Simulation under the Use of Content and Exposure Control ..… 130

LIST OF FIGURES

Figure 1. Example of an Item Characteristic Curve (ICC) 17

Figure 2. Item Characteristic Curves of Rasch Model 19

Figure 3. Item Characteristic Curves of Two Parameter Logistic Model 21

Figure 4. Item Characteristic Curves of Three Parameter Logistic Model 23

Figure 5. Distribution of Exposure Rates in CAT Simulation 80

Figure 6. Effect of Item Exposure Rates Before and After Exposure Control 82

LIST OF ABBREVIATIONS

1PLM	Rasch Model
2PLM	Two Parameter Logistic Model
3PLM	Three Parameter Logistic Model
CAT	Computerized Adaptive Testing
CTT	Classical Test Theory
ICC	Item Characteristic Curve
IRT	Item Response Theory
MLE	Maximum Likelihood Estimation
OPLM	One Parameter Logistic Model
PMS	Pupil Monitoring System
PP	Paper and Pencil Test
SE	Standard Error

CHAPTER 1

INTRODUCTION

The general aim of testing is mapping of examinee ability estimations on the same scale so that the test results can be used for different purposes such as measuring prerequisite skills, monitoring learning process, identifying learning difficulties, assigning grades and providing feedback. Until recently, linear tests, which contain a set of predefined items, have been using classical test theory (CTT) framework. However, the development of new and more psychometrically sophisticated theory, named as item response theory (IRT), provides a well-founded theoretical framework for estimating ability levels of individuals, linking and equating measurements, evaluating test bias, differential item functioning and adaptive testing (Glas, 2010; Scheerens, Glas & Thomas, 2003; Kolen & Brennan, 1995; Hambleton, Swaminathan & Rogers, 1991). Today, linear tests based on item response theory framework are administered all around the world either as a paper and pencil (PP) test or as a computer-based test (CBT) for student placement, professional licensure and monitoring student development. For instance, both PP and CBT versions of Test of English as a Foreign Language (TOEFL) have been delivered in 88 countries.

Different than CTT, design and implementation of a linear test using IRT framework behind requires a list of necessary steps to be followed. Van der Linden and Pashley (2010) point out these steps as follows. First of all, test specialists subjectively rate the difficulty of the newly written items. After pre-

testing these items, data is analyzed by using item response theory framework to obtain the surviving items. Then, a group of items are selected to create the final test forms. These fixed linear test forms, containing anchor items, are checked, modified and administered to a large number of examinees. Psychometricians analyze test results and place the number of correct items for each of the test forms on a common scale by using IRT scaling and true-score equating. On the other side, the examinees have to wait for a great amount of time for the results. To summarize, a great deal of manpower and a plenty of time is needed even on item selection (formation of the test forms), ability estimation (scaling and equating the scores by using item response theory models) and reporting the results.

Instead of administering a fixed set of items as a linear test, there is another option that also uses IRT framework but directs the items according to the ability levels of the examinees. This is the main idea behind adaptive (or in some resources it is called as tailored) testing in which an algorithm matches the difficulty of items with the performance of the examinees during test administration. That is to say, high-ability examinees face with relatively more difficult items whereas low-ability examinees meet with the easier ones (Kingsbury & Hauser, 2004; Bergstrom, Lunz & Gershon, 1992; Hambleton, Swaminathan & Rogers, 1991).

Adaptive testing can be designed to obtain maximum information about the examinees if their ability values are known; but if they are known, then there is no need for testing the examinees anymore. This problem is defined as *paradox of test design* in the literature. Mostly, the solutions to this paradox are generally focused on calibrated item banks which contain a set of well defined items. Additionally, previous responses of examinees are used for item selection and current ability estimations. However, it is not feasible for delivery systems to calculate ability estimations and quickly select items from the bank in the past. Lord (1980) offers two other remedies. First solution is the use of two-stage testing (a routing test is given at first and after quickly scoring it manually, main test is administered to determine ability estimations). Moreover, second one is the

flexi-level testing (examinees score their own responses and if the response is wrong the test directs the examinee to an easier item) (Van der Linden, 1995). Actually these remedies provide good indications that adaptive testing really needs the development of computational technology.

1.1 Computerized Adaptive Testing: Past and Present

Although the idea of adaptive (tailored) testing has roots from the practice of oral examinations (because an expert oral examiner has an impression about the examinee's knowledge level and knows how to tailor the following questions accordingly), the first written exam satisfying the basic characteristics of adaptive testing is Binet & Simon intelligence test (as cited in Weiss, 2004). It is, actually, as early as 1905 when they prepare a paper and pencil based intelligence test to evaluate the mental age of the children. Items are classified according to the mental age. The mental ages of students are calculated from their responses to earlier items and this process continues until predefined sufficient certainty condition is satisfied (Van der Linden; 2010).

Although few studies have been conducted about adaptive testing in the first half of the 20th century, development of item response theory (IRT) framework in 1950s provides modeling of response probabilities in terms of item parameters and ability estimations. Generally, researches in 1960s are condensed about estimation methods and alternative adaptive formats of linear tests. However, the studies of Birnbaum (1969) cause a gradual shift from using classical test theory (CTT) to item response theory (IRT). Furthermore, Frederic Lord (1980) from Educational Testing Service (ETS) makes important contributions to adaptive testing literature in 1970s. Personnel Management of US Armed Service supports studies related to adaptive testing but there are some important barriers on administering computerized adaptive tests in these years, which is obviously insufficient computer speed and high cost of adaptive test implementation (Hambleton et al., 1991).

After computers become available and support high-level computing in late 1980s, some institutions work on developing and implementing a computerized adaptive test. Especially in military, U.S. Department of Defense realizes the benefits of adaptive testing and develops the computerized adaptive test of Armed Services Vocational Aptitude Battery (CAT-ASVAB). Also in health, National Council of State Boards of Nursing (NCLEX/CAT) put the computerized adaptive form of the licensing exam into practice. Furthermore, ETS develops CAT version of General Record Examination (GRE) (Van der Linden & Glas, 2010). In 1990s, Graduate Management Admission Council (GMAC) conducts several studies on the administration of Graduate Management Admission Test (GMAT) as a computerized adaptive test (Rudner, 2010). Computerized adaptive testing also becomes popular in Europe. For instance, the National Institute for Educational Measurement (Cito) in the Netherlands has studied on the development of computerized adaptive tests related to placement and achievement testing of mathematics courses in adult basic education. For this purpose, MATHCAT testing system has been developed and used in Dutch colleges for basic adult education since 1999 (Verschoor & Straetmans, 2010).

So what are the underlying reasons that these institutions have developed systems and preferred the implementation of computerized adaptive testing? At this point, the advantages and limitations of CAT are discussed in order to provide an insight about the CAT administration.

1.2 Advantages and Limitations of Computerized Adaptive Testing

1.2.1 Advantages of CAT

The advantages of computerized adaptive testing over linear tests are grouped under 3 circumstances:

A. Administrational issues

According to Verschoor & Straetmans (2010), Eggen (2007), Wainer (2000), Meijer & Nering (1999) and Bergstrom, Lunz & Gershon (1992), CAT substantially shortens test length and obtains accurate ability estimations with nearly half of the average number of items needed for the linear tests. Likewise, Van der Linden (1995) agrees that CAT reduces the test length and spent time; also shares that computerized adaptive version roughly uses 40 percent of the items and 50 percent of the time compared with its original linear test. In another study conducted by Eggen and Straetmans (2000), reduction rate of the items in CAT is between 22% and 44% compared with the items used in linear versions.

As an additional advantage, instead of administering a test on a strictly fixed date, CAT allows students to schedule test date according to their calendars (Glas & Geerlings, 2009; Van der Linden, 2001; Wainer, 2000; ETS, 1994).

Furthermore, the issue of test security has a great importance in test administrations. In a CAT algorithm, large item bank helps to control cheating of examinees since each examinee take different set of items in a test. Moreover, Wainer (2000) shares the following analogy to support the enhanced test security of CAT: “a test is safer in a computer than it is in a drawer of the desk”.

B. Assessment and scoring

Three advantages of CAT related to assessment and scoring issues are given as follows. First of all, CAT enables immediate test scoring and reporting (Eggen, 2007; Wainer, 2000; Meijer & Nering, 1999; ETS 1994; Wainer, 1993;

Hambleton et al., 1991). Second, CAT allows pretesting of newly written items within real test administrations (Wainer, 2000). Finally, CAT enables enhanced measurement precision of scores (Glas & Geerlings, 2009; Meijer & Nering, 1999; Weiss, 1982).

C. Testing environment

Computerized adaptive testing is also advantageous for its environmental characteristics. For instance, CAT enables integration with multimedia environments and facilitates different kinds of item formats (Glas & Geerlings, 2009; Wainer, 2000; ETS, 1994). Building such a multimedia-based testing environment allows test takers to manipulate the objects appearing on the screen or to work with application programs within the computer (Hambleton et al., 1991). Moreover, instead of taking a test with a large group of students, CAT provides a more comfortable setting with few people (ETS, 1994).

1.2.2 Limitations of CAT

Despite its advantages, Wainer (1993) indicates some restrictions of converting a linearly administered test to an adaptive format.

First of all, it is not possible to review and modify responses to earlier items in a computerized adaptive test. However, results of an empirical study conducted by Lilley and Barker (2004) states that modifying previously entered responses has little effect (no significant different in performance for one group and relatively small significant difference for the other group of learners) on final performance.

Furthermore, omission of items is not allowed in a CAT administration so it is not possible to pass the item or leave it blank. Therefore, either these items are treated as incorrect which is resulted as lower scores, or examinees are desired to select one of the options which increases knowing by chance. Replacing the current item

with a new one with the same psychometric characteristics may be a solution for this restraint but overexposure rates of the items needs to be considered in such situation.

Eggen (2004) shares two restraints of CAT within different perspective. First, CAT needs an automated computer environment containing both hardware and software to deliver the tests according to the standards. CAT also supposes the availability of directly accessible, complete and large item bank in which psychometric characteristics of the items are well defined. Under these circumstances, Meijer & Nering (1999) underline cost of CAT because a great amount of financial and human resource is needed to create the computer system, develop algorithm and construct the items and update the item bank. However, once a CAT system is built, unit cost of implementing a computerized adaptive test is far less than a linear test. Second, although many researchers agree upon the property of facilitating different item formats as an advantage of CAT, Eggen (2004) points out that it may also be a limitation since automatic scoring of CAT does not allow the examinee to give free responses to the items. Therefore, most of the CAT algorithms are limited to dichotomously coded (1 for correct, 0 for incorrect) item responses because of practical difficulty of scoring polytomous items automatically during test administrations.

Eggen & Verschoor (2006) denote that CAT may have possible negative side effects (e.g., enhanced test anxiety) on the examinees because psychometrically the optimal item selection is conducted in CAT with the consideration of 50% probability of answering correctly. On the other hand, within the paper and pencil (PP) tests constructed in primary and secondary education, an average student has a higher probability (60% or 70%) of answering the items correctly. Therefore, students may feel that a computerized adaptive test is more difficult than a linear test. On the other hand, another study conducted by Powers (2001) investigates that effect of anxiety on test performance in PP based and computerized adaptive tests of GRE. The result of this study denotes no evidence supporting the effect of test anxiety on both test performances.

Despite its limitations, general consensus is that the advantages of computerized adaptive tests outweigh so that CAT have been preferred and used in many large-scale testing programs such as General Record Examination (GRE) and Graduate Management Admission Test (GMAT) in all over the world. At the moment, a question arises about the feasibility of using computerized adaptive testing in longitudinal assessments such as pupil monitoring systems.

1.3 Pupil Monitoring System (PMS)

Moelands (2010), Glas & Geerlings (2009), and Vlug (1997) define pupil monitoring system (PMS) as the continuous evaluation of pupils over several years to monitor their development. PMS not only supports investigating the development of individuals according to national standards but also allows the comparison of individuals within peer groups. As an example, National Institute for Educational Measurement (Cito) in the Netherlands develops one of the well-known PMS in which coherent sets of standardized linear tests are used to assess 4 to 12 aged pupils in arithmetic, language and world orientation subjects. Pupils are given different standardized linear tests at different time and analyses based on IRT framework allow the representation of test scores on the same scale. Results of PMS allow teachers both to make decisions about the progress of students' learning process and to determine the relative position of pupils compared to peer groups. Glas & Geerlings (2009) criticize using of computerized adaptive tests in pupil monitoring systems.

As a final decision, the authors recommend using CAT in PMS because of two main reasons: measurement efficiency and possibility of testing on demand. It is a general fact that pupils from different ability levels have different rate of growths but the tests need to be informative at each ability level. Therefore, adapting the test items to pupils' abilities has a positive impact on measurement efficiency. The other reason is related to flexibility of testing date and time because using

computerized adaptive tests in PMS facilitates examinees to take the test whenever they feel ready.

1.4 Turkish Pupil Monitoring System

Cito Turkiye has developed a similar PMS both to assess Turkish students' achievement and to monitor their academic and social development continuously since 2006. Despite some special characteristics, main idea behind PMS in Turkey and in the Netherlands is the same: monitoring the development of pupils with IRT based linear tests in each semester of the year. Ozgen Tuncer (2008) and Is Guzel, Berberoglu, Demirtasli, Arikan & Ozgen Tuncer (2009) share basic characteristics of Turkish PMS as follows. First of all, Turkish PMS focuses on the evaluation of higher order thinking skills under 5 content areas: mathematics, science and technology, social sciences, life sciences and Turkish language. Second, Turkish PMS uses basic properties of IRT in item calibration and test designs. Psychometricians generate parallel test forms by using the calibrated items in an incomplete test design. Anchor items are embedded to each test forms in order to equate test forms both vertically and horizontally. Different test forms are implemented to both 5-6 years old kinder garden children and pupils from grade 1 to 8. Pupils are asked to evaluate written materials and make inferences based on the information provided within the items in different formats: multiple choice, hot spot or open ended. Items contain either text-based, sound-based, picture-based, even animation-based representations or their multiple combinations. In a long time period (more than a month), students are free on deciding when to take their tests. Approximately two weeks after the completion of tests, results of Turkish PMS are shared with teachers, school principals and parents to give feedback in each of the subject matter areas.

There are some similarities and differences between the characteristics of pupil monitoring systems in Turkey and in the Netherlands. Using computer-based

environment, focusing on the primary education, using standardized tests for longitudinal assessment, giving feedback to teachers and school administrators, monitoring the pupil development over years and using IRT framework for modeling educational measurement are some identical characteristics of two pupil monitoring systems. However, one of the salient differences between them is the use of computerized adaptive testing. Turkish PMS has computer based linear tests in which anchor items are used both within grades and between grades so that equating scores from different test forms and monitoring pupil development continuously become possible. On the other hand, PMS in the Netherlands is one step further and uses computerized adaptive tests and linear tests at the same time. Additionally, providing proficiency descriptions of pupils' scores in standard setting is one of the leading characteristics of Turkish PMS. Karantonis & Sireci (2006) share the properties a mostly preferred standard setting method, Bookmark method. By this method, IRT is used to map items onto a proficiency distribution where cut scores (standards) are set.

1.5 CAT Administration in Pupil Monitoring System

Item selection and ability estimation procedures of computerized adaptive and linear tests are completely different. These procedures occur in real time within CAT but separately taken into account by psychometricians within linear tests. In other words, computers perform the roles of both test specialists and psychometricians in CAT so that amount of time and effort reduce noticeably in test administration and analysis (Van der Linden & Pashley, 2010). Hence, Cito in the Netherlands have developed several computerized adaptive tests in PMS such as MATHCAT (test for diagnosing mathematics deficiencies of college students), DSLcat (test for Dutch as a Second Language), TURCAT (test for Turkish as a Second Language) and KindergartenCAT (test for measuring ordering, language and orientation in time and space abilities of young children) (Verschoor & Straetmans, 2010; Veldkamp, Verschoor & Eggen, 2007). At the moment, a

question arises. Is CAT application appropriate for Turkish Pupil Monitoring System?

1.6 Definition of Terms

Computerized adaptive testing (CAT): An IRT based testing methodology that tailors test items according to the performance of participants on previous items.

Linear testing: A CTT or IRT based testing methodology that contains fixed set of items.

Item Information: It is the reciprocal of the variance which is formulated as a function of model parameters (ability and item parameters).

Standard error (SE): It is the reliability measure of ability estimation. It is inversely proportional with the square root of test information function,

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}.$$

Item calibration: Process of obtaining item parameters.

Model fit: The degree to which the data correspond to IRT model characteristics.

1.7 Purpose of the Study

The present study aims to investigate the applicability of computerized adaptive testing to Turkish Pupil Monitoring System (TPMS). The study is conducted in three phases. In the first phase, live CAT administration is implemented for 6th and 7th grade pupils based on the longitudinal data obtained from IRT based standardized linear PMS mathematics administration. In the second phase, Monte Carlo simulations and in third phase Post-hoc simulations are conducted. The results of different CAT simulation scenarios are compared to determine the optimum CAT algorithm regarding:

1. different ability estimation methods: maximum likelihood estimation (MLE), weighted maximum likelihood (WML) or Expected a Posteriori (EAP)
2. different starting rules (starting with easy, moderate or difficult initial item)
3. using content control, item exposure control, neither or both
4. different termination criteria (fixed test length or fixed test reliability)

Research problems of the study can be grouped as follows:

Phase 1:

a) Does live CAT administration provide comparable ability estimations in proportion to computerized linear PMS mathematics assessments?

b) Does live CAT administration provide same information about students' proficiency level compared to linear PMS mathematics assessments?

Phase 2:

a) Does the use of different ability estimation methods (WML and EAP) in Monte Carlo simulations produce different ability estimations?

b) Do different initial item difficulties (easy, moderate or difficult) in Monte Carlo simulations produce different ability estimations?

c) Do different termination rules (fixed test length or fixed test reliability) in Monte Carlo simulations produce different ability estimations?

Phase 3:

a) Does the use of different ability estimation methods (MLE and WML) in Post-hoc simulations produce different ability estimations?

b) Does the change in the difficulty of first item (start the test with an easy, moderate or difficult item) in Post-hoc simulations produce different ability estimations?

c) Does the use of content control, exposure control or both strategies in Post-hoc simulations produce different ability estimations?

d) Do different termination rules (fixed test length and fixed test reliability) in Post-hoc simulations produce different ability estimations?

e) Do Post-hoc simulations provide same information about the students' proficiency level compared to linear PMS mathematics assessments?

1.8 Significance of the Study

Today, a great deal of testing programs in education, psychology, management, health or even military has been using computerized adaptive tests. Although large-scale tests are widely used in Turkey, such as Student Selection Examination, Student Placement Examination, Foreign Language Examination for Civil Servants, Graduate Studies Entrance Examination conducted by Student Selection and Placement Center (SSPC) and Secondary Schools Entrance Examination administered by Ministry of National Education (MONE), SSPC and MONE have been using linear tests based on CTT framework including the same multiple choice items in different order (with different booklets). Whether or not you feel ready for the test, there is no option to take at any other time. Therefore, most pupils feel depressed and test anxiety becomes a major problem in this option. Alternatively, computerized adaptive test administrations can be used instead of linear tests because of providing measurement efficiency, shortening test length, supplying immediate scoring, using different item formats, enhancing test security and having a flexible date of test administration advantages.

Cito Turkiye has been using IRT based linear tests on a computer environment to monitor students' development in pupil monitoring system. The design, analysis and implementation of these tests need much time and effort so computerized adaptive tests seem to be the best alternative because of real time item selection methods and ability estimation procedures.

The present study, which is focused on Turkish pupil monitoring system's mathematics assessment data, probably (1) makes contribution to the studies related to the applicability of computerized adaptive testing over linear tests, (2) has a positive impact on the studies developing CAT algorithm via Monte Carlo and Post-hoc simulations (3) provides an insight to policy makers about the alternative test formats and their use in large scale testing programs.

Within the present study, appropriateness of mathematics items to computerized adaptive testing is under concern, it is expected that the results of the current study can be generalized to other content areas such as science and technology, social sciences, life sciences and Turkish literature.

Finally, there are three testing sessions for each grade level in Turkish PMS. For instance, each student is given items from each sub-domain items under three test administrations. Thus, using CAT in PMS will be more efficient in terms of testing time and number of sessions if the results provide similar estimations.

CHAPTER 2

LITERATURE REVIEW

This chapter is mainly focused on the related studies about item response theory and computerized adaptive testing.

Until technological innovations in computers have supported the use of different test forms in educational measurement, researches on testing generally focus on linear tests using classical test theory (CTT) framework. However, the item parameters depend on examinee characteristics and test scores are dependent to particular set of items in CTT. This framework also does not provide obtaining equal measure of precision for each test score and incapable of making predictions about the future performance of examinees. An alternative test theory developed to complement these missing features is known as item response theory.

2.1 Basics of Item Response Theory (IRT)

Item response theory (IRT) has not only been a theoretical basis for tests but also provided much more useful and generalizable information to the test developers (Schultz & Whitney, 2005). The main idea behind IRT is that the probability of a correct response to an item can be written as a mathematical function of examinee and item characteristics. This S-shaped monotonically increasing function, called item response function (IRF) or item characteristic curve (ICC), provides an

opportunity to predict the performance of an examinee by a set of factors, called traits, or abilities (e.g. mathematics ability) and item parameters. The graph of an ICC is given in Figure 1.

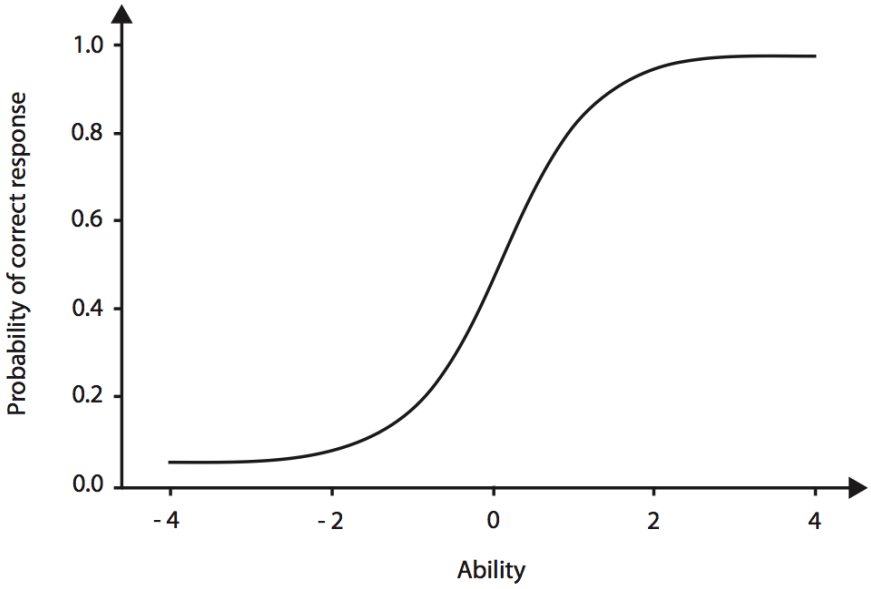

Figure 1. Example of an Item Characteristic Curve (ICC)

In Figure 1, horizontal axis is the ability continuum and vertical axis is the probability of a correct response for a specific item. The graph of ICC indicates that examinees having higher ability values are more likely to give a correct response than the examinees with lower values on the trait.

Unlike CTT, IRT uses probabilistic models to calibrate the item parameters and ability estimations on a common scale. Mainly, there are three basic

unidimensional IRT models for dichotomously coded data: one-parameter logistic model (a.k.a. Rasch model), two-parameter logistic model (2PLM) and three-parameter logistic model (3PLM). In fact, there exists IRT models for polytomous items and multidimensional IRT models but they are out of scope of this research. Unidimensional IRT models have been using either the normal o-give function or the logistic function in ICC. Normal o-give models use the probability of mass under the standard density function but the logistic models are just parametric functions and generally preferred within the studies mostly because of their ease of calculation. A scale factor D (which is equal to 1.7) is used to make the conversion between these two models (Camilli, 1994; Haley, 1952).

2.1.1 Rasch Model

Simplest model of IRT, in which each examinee is represented by an ability parameter and each item is defined by a unique item difficulty parameter, is known as Rasch model (Rasch, 1960). In Rasch model, the number-correct scores of examinees and the number of correct responses given to the items are sufficient statistics for unidimensional ability parameter *theta (symbolized as θ)* and unidimensional item parameter *difficulty (symbolized as b)*. This implies that the examinees' raw scores and number of correct responses to the items contains all the information for the ability estimation and item difficulty parameters. Moreover, probabilities of a correct response as a function of theta are continuous with upper limit 1 and lower limit 0 (Scheerens, Glas & Thomas, 2003).

Rasch model is formulated as:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

where

θ *is the ability value of the examinee,*

$P_i(\theta)$ *is the probability of examinee's correct response to item i,*

b_i *is the difficulty of item i.*

Figure 2 presents three examples of ICCs in Rasch model having different item difficulty parameters *(b=-1, b=0 and b=1)*.

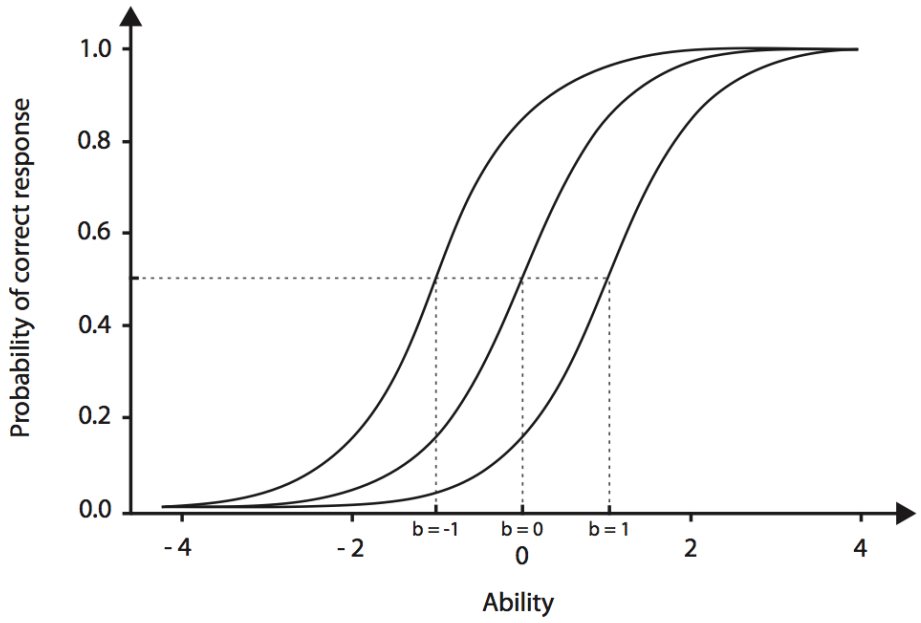

Figure 2. Item Characteristic Curves of Rasch Model

Item characteristic curves presented in Figure 2 are shifted and do not cross each other providing the fact that the order of probabilities of a correct response is the same for all ability levels in Rasch model. Item difficulty parameter *(b)* is equal to the point on ability scale on which the probability of giving a correct response is 50%.

2.1.2 Two Parameter Logistic Model (2PLM)

In some applications of IRT, unfortunately, Rasch model may not provide a good fit to the data. At this moment, a researcher has two options open. One of them is the deletion of the items whose ICCs have divergent slopes and continue on using Rasch model. The second option is using a model containing additional item parameter so that the ICCs have different slopes and intersect each other. For the second option, Birnbaum (1968) defines 2PLM with an assumption that the weighted sum score is a sufficient statistic for ability and the correct response scores are weighted with a discrimination variable symbolized as *a* in the formula given below.

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

where

a_i *is the discrimination parameter of item i,*

D is the scaling factor having a value of 1.7

By setting *a=1* in the above formula gives nothing but the item response function (IRF) of Rasch model. Three different ICCs of items having different discrimination parameters but the same difficulty values are given in Figure 3.

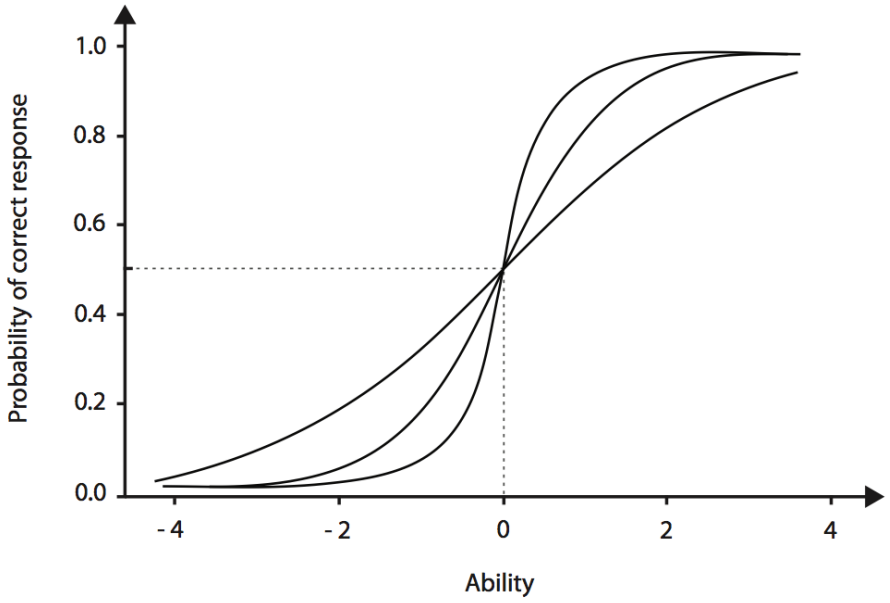

Figure 3. Item Characteristic Curves of Two Parameter Logistic Model

In 2PLM, item discrimination parameter *(a)* characterizes the slope of ICC and indexes the dependence of item response on latent variable theta. For larger values of item discrimination, graph of ICC becomes steeper.

In the present study, 2PLM is used because the data fits to 2PLM. Therefore, item discrimination and item difficulty parameters are obtained for further analysis of computerized adaptive testing during item calibration.

In the literature, there exists an adaptive version of Rasch model, named one-parameter logistic model (OPLM), by Verhelst & Glas (1995) that combines the appealing properties of Rasch model with flexibility advantage of 2PLM. Rather than estimating discrimination parameters as in two-parameter logistic model, OPLM contains preset discrimination indices having integer values between 1 and 15. Item parameters in Turkish PMS are calibrated by OPLM and all the analysis are also conducted by using this model.

2.1.3 Three Parameter Logistic Model

In three-parameter logistic model (3PLM) item difficulty and item discrimination parameters are still remaining but extra pseudo-chance level (guessing) parameter labeled as *c* is entered to the item characteristic curve. This parameter represents the low ability examinees of answering the item by chance and provides a possibly non-zero lower asymptote for ICC.

Three item characteristic curves having different guessing values *(c=0.05, c=0.10 and c=0.20)* are given in Figure 4.

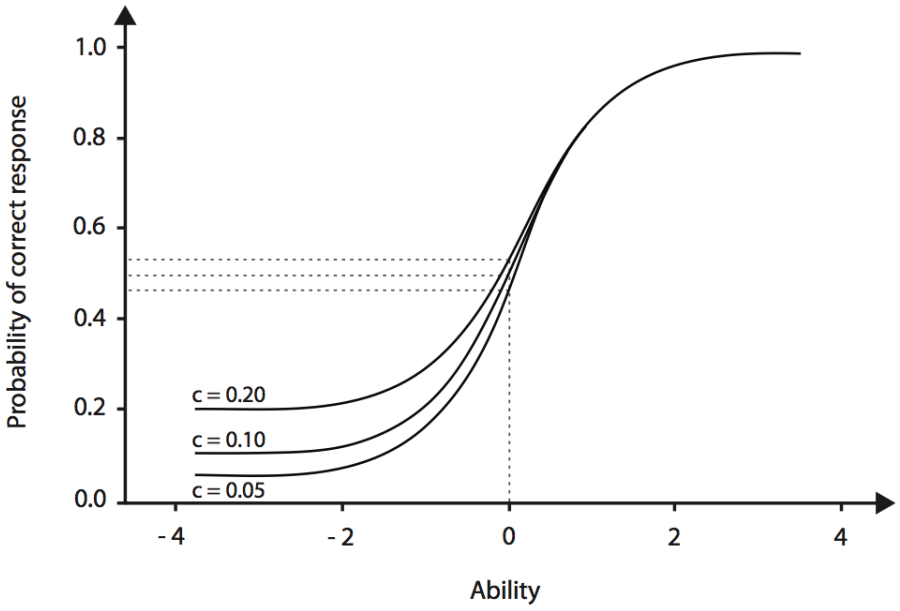

Figure 4. Item Characteristic Curves of Three Parameter Logistic Model

ICC representing 3PLM is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

2.1.4 Information Functions and Standard Error

Reliability is defined as the degree to which a test is free from error. This concept basically involves the notion of ranking the performance of examinees on a test and then re-ranks them on another form of the same test. If the examinees maintain the same order in both rankings, then it reflects the reliability of test (Wainer et. al., 1990). Ranking may be useful and informative but if the performances of the examinees are almost the same, then the test may show a low reliable value. Another major contribution of IRT is the extension of classical reliability (precision of measurement). Rather than having one reliability measure for a test that indicates the amount of true and error variance in observed scores, each item provides an information value about the examinee and contributes to the reduction of uncertainty in examinee's ability estimation. Hence, each ability estimation is denoted with a standard error (SE) value in $\theta \pm SE$ form. Item information function (IIF), denoted by $I_i(\theta)$, is represented by the following formula:

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}$$

where

$I_i(\theta)$ *is the information provided by item i at* θ,

$P_i'(\theta)$ *is the first derivative of* $P_i(\theta)$ *with respect to* θ,

$Q_i(\theta)$ *is the probability function of obtaining an incorrect response to the item i.*

Adapted version of item information function to 2PLM that is used within the study is given as follows:

$$I_i(\theta) = \frac{2.89a_i^2}{\left[e^{1.7a_i(\theta-b_i)}\right]\left[1+e^{-1.7a_i(\theta-b_i)}\right]^2}$$

The above formula exposes the dependency of item information function to item parameters and ability estimations. For instance, information value is directly proportional to the square of item discrimination parameter. Moreover, item information value is maximized when item difficulty approaches to ability estimate. The sum of all item information functions within a test constitutes the test information function.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

This important property of IRT denotes that the administration of each item to an examinee provides a contribution to the test information function and standard error (SE) of examinee's ability estimation $\hat{\theta}$ is inversely related to that amount provided from the test.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I_i(\theta)}}$$

As an example, let an examinee takes 20 items in a test. If the sum of item information values is equal to 25, then the standard error of the ability estimation will be 0.20. In IRT, the relationship between reliability (ρ) and the standard error of ability estimation (SE) is expressed as $\rho=1-SE^2$. The formula implies that 0.84, 0.90, 0.91 and 0.96 reliability values correspond to SE values of 0.40, 0.32, 0.30, and 0.20 respectively.

2.1.5 Model-Data Fit Analysis

In order to use the IRT models for further applications like CAT, a list of model assumptions and expected model features needs to be satisfied.

Unidimensionality: Unidimensionality assumption rests upon the fact that set of items in a test should measure the same unique psychological trait. For instance, a mathematics achievement test that contains items measuring computational ability and some others measuring reading proficiency most probably does not satisfy unidimensionality. Although unidimensionality assumption cannot be strictly met because of several factors affecting the performance like motivation and anxiety in a test, factor analysis results can be used as an evidence for satisfying unidimensionality assumption such that the existence of a dominant factor in eigenvalues provides an indication (Hambleton et al, 1991). Reckase (1979) states that first factor should account for at least twenty percent of the total variance in order to obtain stable item parameters and ability estimates. When an item bank is not unidimensional (then it is multidimensional) there are some alternative steps to do: either separate the item bank into several unidimensional banks or use testlets (a group of items) that contain items from each dimension (Schnipke & Green, 1995).

Local independence: Local independence, which assumes the lack of relationship among responses through latent trait, is closely related to unidimensionality. That

is to say, a correct or incorrect response to an item should not lead to same type of responses to a different item. For example, a reading passage followed by a set of items is a potential source of local item dependence. DeMars (2010) states that if items are locally independent then the items need to be uncorrelated for any values of θ. According to Hambleton et al. (1991), when unidimensionality assumption is met, the local independence assumption is also satisfied.

Equal discrimination indices: According to Hambleton et al. (1991), homogeneous distribution of biserial and point-biserial values indicates an evidence for satisfying equal discrimination indices assumption.

Minimal guessing: The use of multiple-choice items in a test enables finding the answer by guessing. Since there exists a pseudo-guessing parameter in 3PLM, the minimal guessing assumption needs to be checked if the Rasch model or 2PLM is in use. In order to determine the existence of minimal guessing assumption, the performance of low ability examinees on difficulty items needs to be low.

Non-speeded test administration: IRT models assume that each item is seen by the examinees. Therefore, to check the non-speeded test administration assumption, number of omitted responses toward the end of each taskset is under concern.

Invariance of ability estimations: Invariance of ability estimation principle means that each examinee's ability is invariant with respect to the items used to determine it. To satisfy this principle, ability estimates for different samples of test items need to be compared. For example, a set of 20 items having an average item difficulty of -1 is administered to an examinee and ability is estimated as $\hat{\theta}_1$ for this test. Then, another 20 items having an average difficulty of +1 are administered to the same examinee and ability estimation is obtained as $\hat{\theta}_2$. Under the invariance principle, these two different sets of items should be resulted with the same ability estimate ($\hat{\theta}_1 = \hat{\theta}_2$) within sampling variation.

Invariance of item parameters: The parameter invariance property means that the item parameters are not dependent upon the ability level of the examinees. For instance, select a high-ability group and a low-ability group of examinees from a population. If item parameter estimates obtained from these two subgroups do not differ, then this assumption is regarded as met.

Misfit and residual analysis: Scheerens et al. (2003) define the indication of a model fit as the distance between the responses and the expectations. If the expectations are closer to the observation values, then this implies a good model fit. An IRT model is valid if and only if the model fits the data on two perspectives: the items and the respondents. For each item an item-fit statistic is calculated to observe whether the item fits the model or not. If an item fits the model, then item parameter estimate of that item obtained in the different groups of examinees will be the same (with a standard error of measurement). However, if an item does not fit to the model, it will possibly be removed from the analysis. Similarly, for each examinee a person-fit statistic is computed whether the examinee responds according to the estimations of the model or not. When a person fits the model then the ability estimations obtained from different sets of items will be the same (with a standard error of measurement). For instance, if an examinee is not motivated and give a lot of guessed responses, this examinee will be identified in person-fit analysis. Residual analysis deals with the accuracy of model predictions with data. The term "residual" means the difference between observed proportion and predicted probability.

2.2 Studies Related to IRT

IRT provides an opportunity to use partly overlapping tests so that different set of items is administered to different group of examinees. Without the use of anchor items it is not feasible to implement a large number of items to every examinee.

After model-data fit analysis, item calibration, which is the process of estimating item parameters and ability parameter from the data, is conducted regarding the model specifications. In item calibration, examinees are mapped on the same scale with item parameters according to the responses (Veerkamp, 1996). There exist several computer programs that implement item calibration by using different estimation methods such as: LOGIST (Wingersky, 1983), PARSCALE (Muraki & Bock, 1993), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MIRT (Glas, 2010), PARDUX (Burket, 1996) and OPLM (Verhelst, Glas, & Verstralen, 1995). These calibration programs use either conditional maximum likelihood (CML) or marginal maximum likelihood (MML) estimation methods in item calibration. Although both of them aim to find item parameter value that has the highest value for a likelihood function, the main difference between them is the way of handling ability parameter. In one of his study, Eggen (2004, pp 61-96) compares the efficiency of CML and MML estimation in different incomplete designs. The results of this study show that both CML and MML perform almost equally well in all studied designs. Glas & Geerlings (2009) and Scheerens et al. (2003) indicate that it is not feasible to use CML in 2PLM and 3PLM because of the lack of sufficient statistics assumption. Hence, the use of MML estimation method is offered in 2PLM and 3PLM but MML assumes the random distribution of θ in the population. Eggen and Verhelst (2011) state that BILOG-MG uses the MML estimation method in Rasch model, 2PLM and 3PLM. Regardless of the method used, when an item or a group of items violate the model during calibration, then these items are left out of the process and the surviving items are stored in the item bank for the use of computerized adaptive testing.

Studies in the literature about IRT mainly focus on investigating the assumptions, determining the model-data fit and carrying out different item calibration methods. For example, Tang and Eignor (1997) investigate the calibration of dichotomously and polytomously scored TOEFL items. TOEFL Vocabulary and Reading Comprehension and Test of Written English (TWE), and TOEFL Listening Comprehension and Test of Spoken English (TSE) are calibrated using

a combination of 3PLM for dichotomously scored items, Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM) for polytomously scored items. In order to check the unidimensionality assumption, a principal component analysis is conducted and the factor analysis results demonstrate that the percentage of variance accounted by the first component is greater than 40% in all cases. For item calibration, PARSCALE computer program is used and Chi-square fit statistics are calibrated for all 105 items in three TOEFL/TWE combined forms and all 96 items in the two TOEFL/TSE combined forms. Only a small number of polytomously scored items on TOEFL/TSE appear to be poorly fit by the models. Results of the study indicate that the data of TOEFL Vocabulary and Reading Comprehension sections and TWE are reasonably well fit by a unidimensional IRT model with items characterized by a combination of 3PLM&GPCM or 3PLM&GRM. Similarly, the data of TOEFL listening comprehension with TSE indicates well fit by the same combination of models.

In another study conducted in Turkey, Onder (2007) investigates model-data fit statistics of OZDEBIR OSS 2004 science test data containing 45 items to IRT models. The study aims to determine which IRT model provides best fit to the data. The results of the factor analysis indicate that the dominant factor explains the 22% of the total variance explained. The percentages of correct responses to the most difficult 5 items by low ability students are analyzed to determine the minimal guessing assumption. Additionally, omitted responses on first and last five items are analyzed to check the non-speeded test administration assumption but the results indicate that percent missing values of last 5 items are much higher than the values of first 5 items. Therefore, non-speeded test administrations assumption is not met. Invariance of item statistics and invariance of ability parameter model features are tested within the study. Misfit analysis of the study reveals that the best fit is obtained with 2PLM and 3PLM.

Akyildiz (2003) compares scores of examinees taking Student Selection Examination ability estimation with ability estimations obtained from 3PLM of

IRT. The correlations are statistically significant and lie between from 0.47 to 0.60 for different subtests.

The study conducted by Ercikan, Schwarz, Julian, Burket, Weber & Link (1998) focuses on whether multiple choice (MC) and constructed response (CR) (e.g. performance assessment, open-ended, short answer) items can be combined to create a common scale. Different tests containing both MC and CR items in reading, language, mathematics and science subjects of 3rd, 5th and 8th grades are designed to assess the outcomes. The tests are administered to randomly assigned equivalent groups of 800 examinees. By using PARDUX, MC items are calibrated with 3PLM and CR items are calibrated with two-parameter partial credit (2PPC) model. Results of the study indicate that MC and CR items can be calibrated on a common scale. Moreover, an increase in the test length by combining two item types has a positive impact on measurement accuracy. In some parts of the ability scale, adding more MC items does not provide enough information so the use of CR items especially for very low ability and high ability examinees gives better measurement accuracy.

2.3 Basics of Computerized Adaptive Testing (CAT)

Computerized adaptive testing (CAT) is an assessment process in which the test is designed according to the previous responses of examinees. CAT algorithm is responsible for matching the difficulties of items with ability levels of examinees until the termination rule is satisfied (Mills & Stocking, 1996). For a better match of items with examinees, a large item bank containing items form a broad range of difficulties and discrimination is needed.

Van der Linden (1995) lists the following 4 steps of developing an iterative CAT algorithm (algorithm is the set of rules that determines how CAT starts, continues and terminates).

1. *Choosing the ability estimation method*
2. *Deciding on the item selection criteria*
3. *Defining the starting rule*
4. *Determining the termination rule*

1. Choosing the ability estimation method: The first step of CAT algorithm is to determine the ability estimation method. Hoijtink & Boomsma (1995) discuss four ability estimation methods that are mostly used in the literature: maximum likelihood estimation (MLE), weighted likelihood estimation (WLE), expected a posteriori estimation (EAP) and maximum a posteriori estimation (MAP). MLE mainly concentrates on maximizing the likelihood function (symbolized as $L(\theta)$) over the range of possible values of θ. Maximum value of likelihood function is found by taking the first order derivative with respect to θ and solve it by equating to zero. Additionally, the variance of ability estimation is found by two-step operation; first taking additional inverse of the second order derivative and then inserting the estimated θ (Scheerens et al., 2003). However, there is an important thing to be mentioned here is that MLE does not provide estimations for response patterns having all items correct or all incorrect (Glas, 2010; Embretson & Reise, 2000) and MLE should not be preferred in small number of test items (Hambleton et al., 1991).

Warm (1989) states that instead of estimating ability according to the mode of likelihood function as MLE, ability estimations should be based on the mean of likelihood function and offered the use of weighted likelihood estimation (WLE). Although, MLE and WLE give similar results, Warm points out that WLE is less

biased than MLE. Both MLE and WLE use the square root of reciprocal of test information as SE.

Other ability estimation methods, expected a posteriori (EAP) and maximum a posteriori (MAP) are Bayesian ability estimators that are the mean and mode of the posterior distribution, respectively (Veerkamp, 2006). By using the Bayes' theorem a posterior density function $\pi(\theta;x,u)$ of θ is calculated as:

$$\pi(\theta;x,u) = \frac{p(\theta)L(\theta;x,u)}{\int_{-\infty}^{\infty} p(\theta)L(\theta;x,u)d\theta}$$

where

$p(\theta)$ is the prior density function reflecting prior information about θ.

According to Gu and Reckase (2007), Bayesian ability estimators produce small standard errors than MLE for the same number of items but may produce biased estimation when inappropriate prior distributions are selected. Both EAP and MAP use standard deviation of posterior distribution as SE.

2. Deciding on the item selection criteria: Selection and implementation of maximally informative items is the main criteria of item selection procedures in CAT so that high-ability examinees do not face with easy items and low-ability examinees do not need to be administered difficult items because such items provide little information about the examinee's ability (Hambleton et al., 1991). Kingsbury and Hauser (2004) point out that the item selection procedures are designed to provide as much contribution as possible to the precision of ability

estimation. Therefore, the items existed in the bank needs to have high quality; that is to say, the items are piloted, content classification is well defined, characteristics and psychometric features need to be stated properly. Eggen (2007) defines item bank (some resources call it as item pool) as the large collection of items, or reservoir of items that is constructed to measure a well-defined area of knowledge or expertise.

Two main approaches of item selection in CAT are the maximum information criterion and Bayesian item selection criteria. Wainer (2000) states that both item selection procedures give good results but according to Eggen (2004) Bayesian item selection criteria need more demand on computer capabilities. Therefore, maximum information at the current ability estimate is used as an item selection criteria within the present study but details of Bayesian item selection criteria can be found in Van der Linden & Pashley (2010).

3. Defining the starting rule: Starting rule simply considers the selection of first item. Is it better to start with an easier item, or the item one having moderate difficulty or even a difficult item? General notion in the literature is that starting with a difficult item may raise the anxiety of the examinees that is possibly resulted as a negative test performance. However, at the beginning of computerized adaptive test administration, the algorithm does not have any information about the test takers ability. According to Mills and Stocking (1996), many testing programs generally want initial success experience of test takers so they prefer to administer the first item slightly easier than average. On the other hand, there is a potential drawback of this approach that it may be resulted as the overuse of easier items.

The present study defines 3 different starting rules considering the item difficulty of the first item:

$$-1.50 < b < -0.50 \quad (1)$$
$$-0.50 < b < +0.50 \quad (2)$$
$$+0.50 < b < +1.50 \quad (3)$$

Effect of variations in first item's difficulty over ability estimations is analyzed within Monte Carlo and Post-hoc simulations to determine the best fitting starting rule to the data.

4. Determining the termination rule: Researches on the termination of CAT algorithm generally focus on two different termination rules either a *fixed length test* or *fixed test reliability (a. k. a. variable length test).* A *fixed-length test* ends after a specified number of items are administered to each examinee and has a positive effect on the validity. However, measurement precision is not controlled for ability estimations in fixed length tests. Costa, Karino, Moura & Andrade (2009), Dodd, Koch & De Ayala (1989) and Stocking (1987) use fixed test lengths of 10 to 30 items.

Second criterion is the *fixed test reliability* in which the algorithm stops after obtaining precise enough ability estimation but there is a risk of administering too few or too many items to examinees. Essentially, it is guaranteed that all ability estimations are obtained with identical standard error values so reliabilities of all estimations are coherent. General tendency on using fixed test reliability value is 0.90 that corresponds to 0.32 SE.

In the simulation phases of the study, ability estimations obtained from different fixed test lengths containing 15, 25 and 35 items and different fixed test reliability values of 0.84 (for SE=0.40), 0.91 (for SE=0.30) and 0.96 (for SE=0.20) are compared with Turkish PMS mathematics assessment results.

While deciding on the CAT algorithm, two extra constraints, which are *item exposure control* and *content balancing,* need to be taken into account.

Item Exposure Control: In a CAT administration, examinees face with different set of items. Items are selected according to the information value they provide. Hence, some items are administered very frequently and resulted as over exposure of items while some of them are never or hardly ever used and resulted as under exposure of items. Either to overcome the risk of items becoming known or to use items efficiently, different item exposure controls are undertaken by appending some restrictions to item selection algorithm. Sympson & Hetter (SH) (1985) method is one of the most widely used item exposure controls and presents an effective solution based on a probabilistic experiment regarding selection rate of items in a number of simulations. More precisely, item selection by comparing a randomly generated value with selection rates is the main idea behind SH exposure control.

Although Eggen (2001) states that interferences to adaptive algorithm affect the efficiency of CAT, he shares some favorable results about using item exposure control of WISCAT test package conducted in Cito. Without any exposure control 226 out of 680 items in the bank are used which implies 66% of the items are not used. 19 items occur in more than 40% of test administrations. Then, the effect of using SH method for overexposure and progressive method developed by Revuelta & Ponsada (1998) for underexposure are tested. Progressive method is based on the mixture of two criteria: chance and maximum information at current ability estimate. Results of the study show that combined application of these two methods provides better results such that maximum exposure rates are around 0.30 and the number of items used is above 500 out of 680 within any of the simulations.

Georgiadou, Triantafillou & Economides (2007) review item exposure control strategies of CAT developed between the years 1983 and 2005. Researchers group exposure control strategies under 5 classifications as: randomization strategies, conditional selection strategies, stratified strategies, combined strategies and multiple stage adaptive test design.

Item exposure controls are not used in live CAT administration of the present study but Post-hoc simulations investigate the effect of Sympson & Hetter item exposure control strategy (an example of conditional selection) on ability estimations. Detailed information about the other control strategies can be obtained from Georgiadou, Triantafillou & Economides (2007).

Content Balancing: Sometimes measured trait contains sub-domains and examiner demands the representation of these sub domains in a certain proportion in a CAT algorithm. This is either because of providing evidence for content and face validity of the test or because of possible requirement to report separate ability estimations on the sub-domains for diagnostic purposes (Eggen, 2007). This additional constraint to CAT algorithm is named as content balancing (or content control). One of the simplest and most elegant methods is proposed by Kingsbury and Zara (1989; 1991) in which the procedure selects the most informative items from the sub-domains with largest discrepancy between the percentages of items already administered and desired percentages. Within the Post-hoc simulations of the present study, the effect of using Kingsbury & Zara content control is tested.

2.4 Studies Related to CAT

Until the development of microcomputers with enhanced capabilities, computerized adaptive testing (CAT) is economically and practically unfeasible. But as the usage of computer technology in measurement becomes widespread and cost of computing has declined rapidly. Therefore, more studies are conducted in different settings of CAT. As a specific example, almost 25% of all paper presentations at the annual meeting of the National Council on Measurement in Education are related to CAT (Orcutt, 2002). Today, researches in the literature about CAT are mainly focus on the choice of IRT model, characteristics of item bank, starting rule, termination criteria, item selection, ability estimation, exposure control and differential item functioning (DIF).

In the related literature, there are many studies investigating the applicability of a linear test as computerized adaptive. Researchers mainly focus on determining the optimum CAT algorithm and test several ability estimation methods, different starting and termination rules, a variety of item selection methods and the use of content/exposure controls either by live CAT administrations or simulations. For example, a study carried out by Verschoor and Straetmans (2010) investigates the use of computerized adaptive placement test instead of administering a two stage testing in one of the mathematics courses in adult basic education in the Netherlands. In order to place the students to the course offered at three different levels, each student take a routine test of 15 items with an average difficulty in the first stage. According to the scores obtained from this routine test, the examinees take any of the three follow-up tests containing 10 items each. In the study, MATHCAT adaptive testing system is used which delivers tests for both placement (place examinees into courses in three levels) and achievement purposes (monitor the students' achievement during the course). The item bank of MATHCAT contains 578 items covering basic concept and skills, geometry, statistics and algebra. The items are either short-answer, multiple choice or multiple-response formats. 476 of the items are calibrated by using OPLM. In order to estimate the item parameters OPLM software is used. In computerized

adaptive placement testing sessions in order to make the examinees feel comfortable the first two items are selected from easy items. Moreover, the following items are selected by using the maximum information (MI) criterion and the test stops when the examinee is assigned to a course level with 90% certainty, in other words 90% confidence interval of the examinee's current ability estimate no longer covers the cutoff scores of three levels. Under this constraint, the placement tests have different lengths changing from 12 to 25 items. Furthermore, the results of the placement test are compared with paper and pencil version of the test and obtained accurate outcomes that the percentages of correct decisions equal to 88.5% (Level-1), 81.6% (Level-2) and 91.1% (Level-3). The computerized adaptive achievement testing session is somehow different than the placement session in that it is composed of three phases. In the first phase, 10 items (first four of them are mental arithmetic items and the remaining six items are from basic concept and skills) are administered. Depending on the ability estimates the examinees, 20 to 30 items that are administered in the second phase is shaped. In the third phase, 5 pretest items are administered for calibration purpose. After the test the results are reported by graphical representations and short explanatory text. The accuracy of the achievement test is assessed by a simulation study and the results of the first phase are compared with the ones in paper and pencil test. The percentages of correct decisions are 84.8% (Level-1), 76.0% (Level-2) and 87.8% (Level-3). The mean and the standard deviation values of the test length of achievement test are 37.147 and 3.692, respectively. In the conclusion part, the authors share the advantages of using MATHCAT as greater accuracy, shorter test length, less time consuming and easier usage.

Another study carried out by Eggen & Straetmans (2000) is related to the development of CAT algorithm for estimating ability and classifying the examinees into categories. Two different testing algorithms (maximum information (MI) at the current estimate versus administering the item that maximizes the item information at the cutting point of the classification) and five different item selection procedures (1) random (R), (2) maximum information

(MI), (3) maximum information with content control (MI+C), (4) maximum information with exposure control (MI+E) and (5) MI with both content and exposure control (MI+C+E) are investigated. Content of mathematics item bank is composed of three sub-domains: mental arithmetic/estimating (sub-domain A), measuring/geometry (sub-domain B) and other elements of the curriculum (sub-domain C). The item bank is unidimensional and dichotomously coded data fits to OPLM. In the calibration study, 268 mathematics items are administered to a sample of 1,198 students in an incomplete design having 16 different booklets and each of them contains 43 items. Using OPLM computer program, calibration is completed and 250 items (48 from sub-domain A, 49 from sub-domain B and 153 from sub-domain C) fit the model. In order to determine the optimum-testing algorithm for computerized adaptive placement test of mathematics, different cases are investigated but all testing algorithms lead to similar accurate results after the administration of approximately 25 items. Afterwards, the authors are concentrated on the use of item bank. Although, the item bank is used more efficiently (all the items are used in between 5% and 20% of the test administrations) by random (R) item selection procedure, it leads to greatest inaccuracy of ability values. The rest of the item selection methods lead to accurate results; however, they suffer from underuse or overuse of the items. For instance, 132 of the 250 items (52.8%) are never selected and 21 items are used frequently (items occurred more than 20% of the administrations) in MI item selection procedure. Applying MI with content control slightly decreases the number of items never selected (decrease from 132 to 126) but there is an increase in the number of items used frequently (increase from 21 to 32). Using MI with exposure control has a positive effect on the number of items not selected (decrease from 132 to 75) and also decreases the number of items administered frequently (decrease from 21 to 19). The use of MI with content and exposure control has the most positive effect on exposure rates of the items. The number of items that are not selected by the algorithm is decreased from 132 to 53; besides the number of items used frequently is decreased from 21 to 18.

Stocking (1987) conducts two feasibility studies to implement a computerized adaptive test in colleges. First study investigates the comparison of adaptive testing results with conventional tests whereas the second study explores the effect of two termination rules in CAT algorithm: fixed test length and fixed standard error of measurement. Within the first study, the subject of arithmetic is under concern and item bank is comprised of 120 items that are calibrated by LOGIST in 3PLM. CAT algorithm is designed such that maximum information is assigned for item selection and fixed test length with 20 items is set for test termination. Result of the first study indicates CAT is more convenient and offers reasonable improvement on ability estimations. Second simulation study is conducted to determine the optimum termination rule of CAT algorithm for saving testing time and cost. 120 reading comprehension items are used within the study. Fixed length test contains 20 items but a test length limit is set for variable length tests: minimum 10 and maximum 40 items. Results of this study denote that variable-length test is more precise than fixed-length test for low ability examinees. On the other hand, fixed-length tests provided more precise results for high ability examinees. Additionally, obtaining shorter variable-length tests ends up with underestimated true scores for low ability examinees and shorter variable-length tests result in overestimated true scores for high ability examinees.

Kingsbury (2002) conducts an empirical study about the comparison of achievement level estimates obtained from computerized adaptive tests and PP tests. The subject areas are language usage, mathematics and reading. 1200 items are calibrated to a common scale by using Rasch model and maximum likelihood is used within scoring. Correlations between prior PP scores and final CAT scores range from 0.83 to 0.85.

Kingsbury and Hauser (2004) carry out a related study. The two purposes of the conducted study is not only to demonstrate the measurement characteristics of a fixed-form linear test and the adaptive test but also to identify the utilization of these tests for No Child Left Behind (NCLB) project of the US. All the tests and the items are calibrated by using Rasch model. There are four sets of fixed-form

tests across 2 grade levels (grade 4 and grade 8) related to 2 content areas (reading and mathematics). In a reading and mathematics test 40 and 50 items selected respectively according to the grade level content standards. Within the design of the fixed forms, item difficulties are considered as: 36% of the items with difficulties between mean and 1 standard deviation (SD), 9% of the items between 1SD and 2SD finally 5% of the items between 2SD and 3SD. The percentages are also same for standard deviations below the mean. Since item difficulty values are obtained from Rasch model, it allows calculating relevant scores such as scale score (RIT) and standard error of measurement (SEM). In the study, test information values associated with each scores are calculated for both fixed-form tests and for adaptive tests by using the SEM values. On the other hand, since the items administered to each examinee are different in adaptive tests, average of the test information values are taken into account in the analysis. In the result part, the authors mention that the adaptive test provides more information at every level of achievement than either fixed-form test. More specifically, adaptive tests measure at least 99% of the students precisely but this value is at most 94% in either fixed-form test. Also the authors state that it is more appropriate for the fixed-form test to make decisions about categorizing the placement of students but not for making appropriate instructional decisions or measuring students' growth.

In the study of Shermis, Fulkerson & Banta (1996), a computerized adaptive test is developed for placing the fifth grade elementary school children in middle school mathematics talent development program. The item bank contains 240 items but since none of the students would be able to complete a 240 item paper and pencil test, the items are grouped under eight forms and each form has 10 anchor items for equating by IRT. The factor analysis results of the study show that the item bank is essentially unidimensional because the first factor explains 41% of the total variance. In the study, three different samples are used to assess the efficacy of the computerized adaptive math test. The first sample contains 683 sixth-grade middle school children from suburban schools. In the second sample, there are 190 sixth-grade middle school children from the same district but these

children are already placed in a mathematical talent program. Finally, the third sample is comprised of 199 fifth-graders who are the candidates for talent development program in the same district. The eight forms of paper and pencil test are administered to the first and second samples of the study for vertical equating and calibration purposes. Afterwards, the data obtained from these tests are calibrated by using Rascal, an IRT computer program for Rasch model. The remaining 199 fifth-grade talent development candidates are participated in computerized adaptive placement testing. The average of the items administered to each student is 15.6 and the average theta ability estimate of the sample is 0.45 (SD=0.66). Furthermore, the correlations of adaptive math test domain score with other aptitude tests such as Otis-Lennon_Math ($r=0.35$, $p<0.001$) and CTSB_Math ($r=0.44$, $p<0.001$) are also calculated within the study.

Rudner (2010) shares the experience of the Graduate Management Admission Council (GMAC) in implementing a CAT assessment. Graduate Management Admission Test (GMAT) is a standardized test assessing the qualifications of applicants for advanced study in business and management. GMAT contains three main components, Analytical Writing Assessment (AWA), Quantitative section and Verbal section and data fits to 3PLM. In 1996 and 1997, two studies are designed on the comparison of PP and CAT versions of GMAT. PP and CAT test results do not give comparable results at first. In spite of the negative outcomes, researches and discussions over the implementation of CAT are continued. Rudner gives information about the outcomes of the issues and the approach taken by GMAC. First, GMAC separates the specifications for the individual and specifications for the item bank. That is, GMAT Quantitative items are classified into three categories as skill area (data sufficiency or problem solving), content base (algebra, arithmetic skills or geometry) and application (applied or formula based). In a CAT application, each examinee takes a certain number of items from these categories. Second, before the item exposure controls are implemented to GMAT, 28% of the items are never used, 18% of the items in the bank are seen more than 15% of the examinees in the previous studies of GMAT. After the

overexposure of items, a mixture of randomization exposure control strategy is developed to satisfy the distribution of items ideally. Third, GMAC also focuses on the item bank characteristics of CAT. There are at least 9,000 high quality items in the bank of GMAT that cover each content requirement at each ability level for producing a highly satisfactory CAT. Fourth, in a routine basis GMAC also examines the item bias by investigating differential item functioning on the item parameters resulting from group-specific calibrations during pretesting. Finally, GMAC has simultaneously recalibrated the entire item response data including the nonoperational items to investigate the consistency (whether there is a parameter shift) of GMAT item parameters.

CAT is also used in studies related to the enhancement of English as a Second Language (ESL) proficiency. Not surprisingly, these studies end up with similar results. For example, Molina (2009) develops a computerized adaptive vocabulary test regarding the enhancement of ESL proficiency in lexical competence. In the first phase of the study, a paper and pencil vocabulary test containing 220 multiple-choice items is constructed. Each of the items contains 5 answer options plus a last option "none of these" to lessen the guess factor. The words used within the items belong to different lexical word classes (nouns, verbs, adjectives and adverbs) in proportion to their percentages reflected in the ADELEX (Assessing and Developing Lexical Competence through the Internet) Project list, which is composed of 7,000 words selected from different resources such as British National Corpus and Longman Corpus. Therefore, 120 items are nouns, 52 items are verbs, 40 are adjective and 8 are related to adverbs. This paper and pencil based vocabulary test is administered to 330 undergraduates of English Philology and to the students of Translation and Interpretation Studies at two different Spanish Universities: Granada and Jaume-I. In order to obtain more objective and reliable results, paper and pencil vocabulary test is changed into a computer adaptive test. The adaptation process is carried out in three stages: (1) calibration of the item bank according to IRT, (2) comparison of item parameters obtained from CTT and IRT and (3) implementation of CAT. As a CAT

algorithm, Maximum Likelihood (ML) ability estimation and Maximum Information (MI) item selection is used in 3PLM. As a termination rule, the procedure stops when the standard error of ability estimation is below 0.32 or a total of 30 items is asked. There is a maximum time of 30 seconds for each item. The results of this study point out that Pearson correlation coefficient between the PP and CAT ability estimation scores of students is 0.94, which leads to a high correlation and offers the use of CAT because of not only the reduction in test length and time but also the precision and reliability of the ability estimation.

Imai, Nakamura, Akagi, Honda, Ito, Kikuchi, Nakasono & Hiramura (2009) states the features of Japanese Computerized Adaptive Test (J-CAT) developed to diagnose the proficiency level of Japanese as a second/foreign language. There is not predefined starting rule of the algorithm. The test administrator can define whether to start with item having medium difficulty or implement few items from a range of difficulties to estimate ability. Similarly, there is not a strict stopping rule. Fixed length or variable length, the test stops whenever the examinee reaches either of the stopping rules defined. Moreover, the algorithm uses maximum information in item selection procedure and Bayesian expected posteriori (EAP) in ability estimation. Maximum likelihood estimation is not preferred because the procedure can not estimate for all correct or all incorrect responses.

Gafni, Cohen, Roded, Baumer & Moshinsky (2009) present the practical issues in designing and evaluating the CAT versions of Psychometric Entrance Test (MIFAM) and English as a Foreign Language Placement Test (AMIRAM). The test contains items from three domains: 210 items from verbal reasoning, 175 items from quantitative reasoning and 230 items from English as a foreign language. Instead of using only a linear test, CAT administration is designed based on 3PLM and the run in parallel with linear test. Evidences about satisfying the unidimensionality assumption is obtained from factor analysis. Moreover, the parameter estimations are obtained by using BILOG-MG. Testing algorithm starts with implementing randomly selected two items from each domain having average level difficulty and low discrimination value. Two termination rules are

defined in the algorithm and the test ends either the examinee reaches the maximum of items defined or the posterior variance is below a predefined value. Exposure controls and content controls are also taken into account in the study. The results of the study show that the scores obtained from both CAT and linear test are similar to each other and have correlation coefficients around 0.80.

Costa, Karino, Moura & Andrade (2009) conduct a simulation study about comparing the performance of three item selection methods for CAT. The maximum information (MI), Kullback-Leibler information (KL) and maximum expected information (MEI) are tested on 246 items calibrated by 3PLM on a (0,1) scale. The data is obtained from Instrumental English I course in University of Brasilia. In the first simulation study, the necessary number of CAT items to obtain ability estimations with a standard error (SE) less than or equal to 0.20 and 0.40 are assessed. The average number of items in 500 simulations to obtain ability estimations is almost the same for three item selection methods and calculated as around 20 items for SE=0.40 and around 140 items for SE=0.20. In the second simulation study, ability estimations of the three item selection methods are compared with a fixed-length test containing 25 items. All three methods estimate θ quite well but they perform worst for $|\theta| \geq 3.50$ because only few items in the bank have such extreme difficulty parameters. Third simulation study is designed to compare the three item selection methods when the initial ability values are set θ=-1.50, θ=0.00 or θ=1.50. For each of the initial ability values, θ is replicated 100 times with fixed length of 25 items to observe the variability. The results of this simulation study show that the procedures for adaptive selection efficiently estimate ability independent of the initial value. Fourth simulation study aims to compare the qualities of ability estimations with different difficulty parameters for the first CAT item. Each of the ability values θ=-1.50, θ=0.00 or θ=1.50 is replicated with 100 times and the test length is fixed to 25 items. Three analyses are conducted with easy, moderate and difficult initial item selections and the results show that MI, KL and MEI item selection methods estimate the ability with the same precision. Therefore, the item selection methods

are independent of the initial item difficulty parameter. Last simulation study is designed to assess the measures bias and mean square error (MSE) with 10 different θ values. Within the 100 replications of θ, the bias and MSE of the estimates are calculated for test lengths 1 to 30. The results show that MSE decreases, as the number of items increase.

Also there exist a few studies related to computerized adaptive testing in Turkey. Koklu (1990) compares validity and reliability of PP and adaptive versions of tests and indicates that adaptive tst administration provides slightly beter results than PP version.

Another study comparing the ability estimations of CAT and PP versions of mathematics test is implemented by Kaptan (1993). The examinees are given a PP test containing 50 items and a computerized adaptive test with 14 items. CAT algorithm uses ML ability estimation method. Results of the study indicate that implementing computerized adaptive test reduces the number of items used in a test by 70%. Moreover, there is no significant difference between the ability estimations obtained from both tests.

Iseri (2002) explores the applicability of IRT to mathematics sub-tests of Secondary Schools Entrance Examinations (SSEE) and Private Schools Entrance Examinations (PSEE). Different ability estimations (MLE and MAP), use of exposure control strategies, test termination rules (fixed test length, fixed test reliability) and allowing revision of previous responses are tested. 187 items from 1998 to 2001 are calibrated with BILOG. Results of the study show that MAP estimation provides better results than MLE. The optimum CAT termination rule is found as fixed test reliability of 0.32 SE. Additionally, the use of exposure controls provides no difference in ability estimations and the correlation coefficients between PP and CAT lie within 0.78-0.92 range.

Lastly, a study conducted by Kalender (2011) compares the ability estimations obtained from both computerized adaptive and PP tests of Student Selection

Examination (SSE) science subtest. In the study, researcher conducts both Post-hoc simulations and live CAT administration to determine the optimum conditions considering different ability estimation methods and test termination rules. Findings of Post-hoc simulations indicate that EAP with fixed test reliability termination rule of SE<0.30 is the optimum CAT algorithm. Moreover, MLE estimation method needs fewer items than EAP to estimate examinees' abilities. Correlation coefficient between ability estimations obtained by CAT and real SSE is found to be 0.95 in Post-hoc simulations. On the other hand, the correlation coefficient between live CAT results and real SSE ability estimations is resulted as 0.74. Average number of items given to examinees in CAT administration is calculated as 18.4.

2.5 Summary of the Literature Review

1. Item calibration is the process of obtaining item parameters. Various computer programs are used for calibrating items such as BILOG-MG (Kalender, 2011; Imai, Nakamura, Akagi, Honda, Ito, Kikuchi, Nakasono & Hiramura, 2009; Gafni, Cohen, Roded, Baumer & Moshinsky, 2009; Iseri, 2002), LOGIST (Stocking, 1987), OPLM (Verschoor&Straetmans, 2010; Eggen & Straetmans, 2000), MIRT etc.

2. For dichotomously coded items, 2PLM and 3PLM are mostly used logistic models in the studies (Kalender, 2011; Rudner, 2010; Molina, 2009; Imai, Nakamura, Akagi, Honda, Ito, Kikuchi, Nakasono & Hiramura, 2009; Costa, Karino, Moura & Andrade, 2009; Iseri, 2002; Stocking , 1987). But there are also studies using Rasch model (Kingsbury and Hauser, 2004; Kingsbury, 2002).

3. CAT algorithm is composed of an ability estimation method, a starting rule, a termination criterion and an item selection strategy (Van der Linden ,1995).

4. An attention needs to be paid to item exposure and content control. Without an item exposure control strategy, it is possible to have an item with 0.50 exposure rate (Eggen, 2001; Eggen & Straetmans; 2000).

5. In some studies fixed test length termination rule is used (Costa, Karino, Moura & Andrade, 2009; Kaptan, 1993; Dodd, Koch & De Ayala, 1989; Stocking, 1987) but some of them use fixed test reliability termination rule (Kalender, 2011; Iseri, 2002) or both (Molina, 2009).

6. In the literature, correlation coefficients between the linear and CAT versions of a test lie between 0.74 and 0.94 (Kalender, 2011; Molina, 2009; Gafni, Cohen, Roded, Baumer & Moshinsky; 2009; Iseri, 2002; Kingsbury, 2002).

CHAPTER 3

METHODOLOGY

In this chapter, methodology of the present study is reported. Basically, the study has three phases. The first phase is related to Live CAT administration which uses multiple choice items of Turkish Pupil Monitoring System (PMS) mathematics assessment. After data analysis and item calibration, a computerized adaptive testing (CAT) algorithm is designed and testing environment is developed. Finally, a live CAT administration is implemented to a group of examinees who are already taken PMS mathematics assessment. The researcher determines how comparable ability estimations are obtained from the live CAT administration with PMS mathematics assessment scores. After implementing the first phase of the study, researcher investigates the effect of including also open-ended items into the item bank. Thereafter, second and third phases of the study are designed to determine the optimum CAT algorithms for both multiple-choice and open-ended items of PMS mathematics assessment. For this purpose, several CAT algorithms containing various (1) starting rules, (2) termination criteria, (3) different ability estimation methods and (4) use of content and exposure control strategies are tested with both Monte Carlo (MC) and Post-hoc simulations. MC simulations are based on the response data generated randomly from item parameters. Although Wang, Pan & Harris (1999) state that MC simulations may not reflect the psychometric characteristics of the examinees, these simulations inform the researcher about the comparison of different ability estimations, various starting and termination rules with the use of item exposure control for the

item bank defined. To identify the detailed psychometric characteristics of ability estimations, Post-hoc simulations are implemented on the real responses of examinees as the third phase of the study. Ability estimations obtained from these Post-hoc simulations allow the comparison of examinees' ability estimations with their prior scores in PMS mathematics assessments.

3.1 Sample of the Study

The current study uses PMS mathematics assessment data obtained from 12,156 students in 4^{th}, 5^{th} and 6^{th} grades between the years 2007 and 2011. The dataset of each year is comprised of the responses of examinees within the tasksets of geometry (GE), measurement (ME), numbers (NU) and probability & statistics (PS). Table 3.1 represents the distribution of items in each content area.

Table 3.1 Distribution of Items Used in Live CAT Administration Regarding Content Area

	N	Percentage
Geometry (GE)	158	31.2%
Measurement (ME)	82	16.2%
Numbers (NU)	212	51.8%
Probability and Statistics (PS)	55	10.8%
TOTAL	507	100.0%

First phase of the study (live CAT administration) concentrates on dichotomously coded multiple-choice items and calibrations are separately conducted for each task set by using BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996). The results of item calibration process indicate that, 507 multiple-choice items fit the model so item discrimination and difficulty parameters are obtained but 11 items are eliminated because they do not fit to the model. Summary statistics of PMS multiple-choice item parameters calibrated by BILOG-MG is given in Table 3.2.

Table 3.2 Summary Statistics for PMS Item Parameters Calibrated by BILOG-MG

N=507					Percentiles		
Estimated parameter	Mean	S.D.	Min.	Max.	25	50	75
Discrimination (a)	0.628	0.242	0.122	1.729	0.459	0.602	0.778
Difficulty (b)	-0.425	1.228	-3.139	3.401	-1.210	-0.589	0.140

Simulation phases of the study are carried out by using the dichotomously coded multiple-choice and open-ended items. MIRT (Glas, 2010) program is used for the item calibration of 733 items since it allows using incomplete test data set in item calibration. 47 of the items do not fit two parameter logistic model so they are eliminated from the item bank. Remaining 686 items (490 items are multiple-choice type and 196 items are open-ended type) are used in simulation phases. Number of items regarding content area and item types is presented in Table 3.3 and summary statistics of the items parameters calibrated by MIRT is given in Table 3.4.

Table 3.3 Number of Items Used in Simulation Phase Regarding Content Area and Item Types

	Multiple-choice	Open-ended	TOTAL
Geometry	156	11	167
Measurement	81	87	168
Numbers	202	72	274
Probability and Statistics	51	26	77
TOTAL	490	196	686

Table 3.4 Summary Statistics for PMS Item Parameters Calibrated by MIRT

N=686					Percentiles		
Estimated parameter	Mean	S.D.	Min.	Max.	25	50	75
Discrimination (a)	0.917	0.363	0.251	2.161	0.664	0.864	1.132
Difficulty (b)	0.126	1.141	-3.015	3.952	-0.622	0.084	0.804

Statistics given in Table 3.2 and Table 3.4 indicates that expansion of item bank with open-ended items for MC and Post-hoc simulations produces an observable change on the distribution of item parameters. Mean of item discrimination parameter increases from 0.628 to 0.917 and similarly mean of item difficulty parameter increases from -0.425 to 0.126. Actually, this may produce a good indication that open-ended items are more difficult to answer than the multiple-choice items. Likewise, addition of open-ended items to the item bank results with higher discrimination parameters than multiple-choice items. Moreover, correlations among item parameters obtained from BILOG-MG, MIRT and OPLM are relatively high as shown in Table 3.5 and Table 3.6. The item parameters can be found in Appendix A.

Table 3.5 Correlation Coefficients of Item Discrimination Parameters obtained from OPLM, BILOG-MG and MIRT

	OPLM	BILOG-MG	MIRT
OPLM	-	0.856	0.794
BILOG-MG	0.856	-	0.816
MIRT	0.794	0.816	-

Table 3.6 Correlation Coefficients of Item Difficulty Parameters obtained from OPLM, BILOG-MG and MIRT

	OPLM	BILOG-MG	MIRT
OPLM	-	0.879	0.830
BILOG-MG	0.879	-	0.918
MIRT	0.830	0.918	-

3.2 Model Data Fit

Turkish Pupil Monitoring System has been using IRT models and by using OPLM, items are calibrated and model-data fit statistics are already determined. The researcher investigates evidences for the assumptions and some expected model features of IRT. First of all, the unidimensionality assumption is said to be

satisfied if there is a sharp drop from the first eigenvalue to the second. Some resources state that the ratio of the first eigenvalue to the second needs to be focused. Instead of a complete test, PMS data contains a set of incomplete data which obtained from different tasksets. Factor analysis results of 20 tasksets in PMS mathematics assessments are given in Table 3.7.

Table 3.7 Eigenvalues of Factor Analysis Results

TASKSET	COMPONENT 1	COMPONENT 2
Taskset 1	7,928	2,704
Taskset 2	9,386	2,366
Taskset 3	9,007	2,545
Taskset 4	8,907	2,600
Taskset 5	5,841	1,813
Taskset 6	6,169	1,718
Taskset 7	5,849	1,734
Taskset 8	6,429	1,686
Taskset 9	6,457	1,822
Taskset 10	5,874	1,724
Taskset 11	6,034	2,098
Taskset 12	6,001	1,791
Taskset 13	7,020	1,656
Taskset 14	7,012	1,685
Taskset 15	7,130	1,637
Taskset 16	7,223	1,683
Taskset 17	7,488	1,878
Taskset 18	7,791	1,709
Taskset 19	8,099	1,762
Taskset 20	7,886	1,683

According to the factor analysis results, sharp drops from 1st component value to the 2nd indicates evidence to unidimensionality of PMS mathematics assessment data. Hambleton et al. (1991) state that when unidimensionality assumption is met, the local independence assumption is also satisfied. "Equal discrimination indices" is an assumption required for 1PLM so analysis on this assumption is discarded.

Minimal guessing assumption is for 1PLM and 2PLM. This assumption can be checked from the performances of low ability students on difficult items. Table 3.8 points out the percentages of correct responses on most difficult items.

Table 3.8 Percentages of Low Ability Students' Correct Responses on 10 Most Difficult Items

ITEMS	PERCENTAGE OF CORRECT
Item 1	3.35%
Item 2	2.63%
Item 3	4.44%
Item 4	0.75%
Item 5	6.15%
Item 6	1.46%
Item 7	3.26%
Item 8	7.95%
Item 9	5.43%
Item 10	6.45%

Low values of correct percentages in Table 3.8 indicate an evidence that minimal guessing assumption is satisfied.

Another assumption, non-speeded test administration assumption can be verified by investigating the percentages of empty responses for the items located in the last part of the test. Actually, PMS assessments are not speed-tests. 2 minutes are given for each Geometry (GE) items and 4 minutes are given to response Measurement (ME), Numbers (NU) and Probability&Statistics (PS) items. Besides, percentages of missing responses in Table 3.9 provide evidence for the assumption of non-speeded test administration.

Table 3.9 Percentages of Missing Responses on Last 5 Items in Tasksets

TASKSET	PERCENTAGE
Taskset 1	1.3%
Taskset 2	0.5%
Taskset 3	2.2%
Taskset 4	1.8%
Taskset 5	3.2%
Taskset 6	2.1%
Taskset 7	1.9%
Taskset 8	3.5%
Taskset 9	1.2%
Taskset 10	2.7%
Taskset 11	2.5%
Taskset 12	2.6%
Taskset 13	3.0%
Taskset 14	1.6%
Taskset 15	2.2%
Taskset 16	1.9%
Taskset 17	0.8%
Taskset 18	3.5%
Taskset 19	3.1%
Taskset 20	2.4%

As it is indicated previously, results of BILOG-MG indicate that 11 of 518 multiple-choice items do not fit the 2PLM in first phase calibrations. However, Lagrange Multiplier (LM) tracelines of MIRT program indicate that 49 of 735

multiple-choice and open-ended items in the second and third phases of the study do not fit to 2PLM.

3.3 Live CAT Administration

The first phase of the study uses PMS infrastructure. After defining the computer algorithm, Atlas Yazilim develops web based live CAT administration program by using .NET framework. 457 students from 4 private and 3 state schools take the administration with the same format and conditions in PMS assessments. Live CAT administration is implemented in July 2011 by using 507 multiple-choice items of 4^{th}, 5^{th} and 6^{th} grades. Basic characteristics of the algorithm used in live CAT administration are stated as follows:

Starting rule: CAT algorithm randomly selects the first item from the set of items having medium difficulty ($-0.1<b<0.1$). There are 31 items satisfying this condition.

Ability estimation: Ability estimation is set as Maximum Likelihood (ML).

Item selection: CAT algorithm selects the items regarding Fisher's Maximum Information.

Termination criteria: CAT administration ends after a fixed reliability value is obtained. For this purpose, standard error is set as 0.30 and the algorithm ends after this condition is satisfied.

Use of item exposure and content controls: No item exposure and content control strategies are used.

3.4 Monte Carlo Simulations

Monte Carlo (MC) simulations are conducted with supervision of Mr. Cees Glas from University of Twente. FORTRAN based computer program developed by Glas are used within this study. Before the simulations, the researcher designs the plan so that two different ability estimations (weighted maximum likelihood or expected a posteriori), three different starting rules (begin to the test with easier, moderate or more difficult item), two different termination criteria (fixed length or fixed test reliability) and use of item exposure control are tested on dichotomous data generated from item parameters. This makes MC simulations allow comparing the true theta obtained from generated responses to all items of examinees with theta estimation gained after a CAT simulation algorithm. Although MC simulations give information about the optimum CAT algorithm for the item bank, it is not meaningful to compare PMS mathematics assessment scores with the scores obtained from MC simulations because MC simulations generate data for randomly selected examinees. On the other hand, Post-hoc simulations which use real data of the examinees allow the comparison of results from PMS assessment and CAT simulations.

3.5 Post-hoc Simulations

The researcher develops web-based simulation software by using PHP programming language and MySQL database for Post-hoc simulations. A set of ability estimations obtained from two different ability estimations (maximum likelihood or weighted maximum likelihood), three different starting rules (three different initial item difficulty ranges are defined as -1.50<b<-0.50 (easy), -0.50<b<0.50 (moderate) and 0.50<b<1.50(difficult)), three different fixed-length termination criteria (with 15, 25 and 35 items), three different fixed test reliability termination criteria (with SE=0.20, SE=0.30 and SE=0.40), use of Sympson & Hetter item exposure control strategy and use of Kingsbury & Zara content

control strategy is tested to determine the optimum CAT algorithm for PMS mathematics data.

Moreover, the researcher also investigates whether the ability estimations obtained from live CAT administration phase and Post-hoc simulation phase provide same information about the students' proficiency levels compared to linear PMS mathematics assessment results. For this purpose, proficiency levels obtained from these two phases are compared with pupils' original PMS proficiency levels.

CHAPTER 4

RESULTS

This chapter presents the results of current study under three titles. First of all, ability estimations obtained from live CAT administration are compared with PMS mathematics assessment scores. Next, the results of Monte Carlo (MC) simulations are indicated. Although MC simulations generate random examinee responses with respect to the item parameters, they produce a great deal of information about the optimum CAT algorithm. Finally, results of Post-hoc CAT simulations dealing with different scenarios on real data are presented in order to determine the optimum algorithm having comparable results with PMS mathematics assessment.

4.1 Results of Live CAT Study

454 students from 6^{th} and 7^{th} grades take live CAT administration. These pupils have already taken the PMS administration so they have 4 sub-domain scores (GE, ME, NU and PS) in mathematics assessments. The researcher evaluates a weighted mean score which is obtained according to the percentage of items in each sub-domain. Since live CAT administration gives a unique ability estimation for each examinee, this value is correlated with each sub-domain score separately and with weighted mean score as well. The correlation coefficients are given in Table 4.1.

Table 4.1 Correlation Coefficients of Ability Estimations between PMS Computer Based Linear Mathematics Assessments and Live CAT Administration

	GE	ME	NU	PS	Weighted Mean
Relationship between Live CAT and PMS	0.65*	0.65*	0.68*	0.66*	0.75*

**All correlations are significant at the 0.01 level.*

Correlation coefficients between live CAT administration and PMS mathematics assessment results lie within 0.65-0.70 range. On the other hand, the correlation coefficient with weighted mean of subdomains is around 0.75 and all the coefficients are significant at the 0.01 level.

Moreover, proficiency estimations across CAT and linear PMS are compared. Ability estimations of Live CAT administration are used to determine the proficiency levels with respect to the cut-off points in PMS mathematics assessment. The percentages of correct proficiency levels in all sub-domains are indicated in Table 4.2.

Table 4.2 Percentages of Correct Proficiency Levels in Live CAT Study

	GE	ME	NU	PS
Correct Proficiency Level	32.5%	64.8%	45.1%	31.7%

The highest percentage of correct proficiency level is in Measurement (ME) items with a value of 64.8%. The other percentages are 45.1% for Numbers (NU) items, 32.5% for Geometry (GE) items and 31.7% for Probability and Statistics (PS) items.

The researcher also investigates the effect of using open-ended items in CAT algorithm so a set of simulations using both multiple-choice and open-ended items are conducted. The results are given as follows. At first, the outcomes of Monte Carlo (MC) simulations are under concern.

4.2 Results of Monte Carlo Simulation Studies

In MC simulations, various starting rules, termination criteria and ability estimation methods are designed and tested to obtain information about optimum CAT algorithm. Simulations are conducted by using both Weighted Maximum Likelihood (WML) and Expected a Posteriori (EAP) ability estimation methods to determine their effect on CAT algorithm.

4.2.1 Determination of Optimum Starting Rule with Fixed Test Reliability Termination Criteria

First of all, 18 different simulations are conducted to test 3 different fixed test reliability termination criteria with 3 different starting rules by 2 different ability estimations: WML and EAP. Average number of items implemented in CAT simulations regarding WML and EAP ability estimations are given in Table 4.3.

Table 4.3 Number of Items Administered under Different Starting Rules and Fixed Test Reliability Termination Rules Regarding WML and EAP Ability Estimations

Estimation	Starting rule		Termination rule: fixed test reliability		
			SE<0.20	SE<0.30	SE<0.40
WML	-1.50 < b < -0.50 (Easy)	Mean	53.83	20.69	11.57
		Min.	38	15	7
		Max.	231	97	38
	-0.50 < b < +0.50 (Moderate)	Mean	54.76	20.19	11.38
		Min.	38	14	7
		Max.	336	73	47
	+0.50 < b < +1.50 (Difficult)	Mean	53.61	21.06	11.44
		Min.	39	15	8
		Max.	289	104	148
EAP	-1.50 < b < -0.50 (Easy)	Mean	69.90	22.19	12.19
		Min.	29	15	7
		Max.	601	163	50
	-0.50 < b < +0.50 (Moderate)	Mean	72.68	21.85	11.68
		Min.	29	14	7
		Max.	579	151	71
	+0.50 < b < +1.50 (Difficult)	Mean	70.14	22.29	12.53
		Min.	30	15	7
		Max.	588	111	86

MC simulation studies developed to test fixed test reliability with different ability estimation methods indicate that EAP needs more items to satisfy the same standard error value for ability estimations. This is favorably more prominent while the 0.20 standard error is used as the termination criterion. Results of these simulation studies also demonstrate that various initial item difficulties of CAT algorithm produces almost same number of items in an administration. Correlation coefficients between expected values and observed ability estimations of WML and EAP methods are given in Table 4.4.

Table 4.4 Correlation Coefficients between Expected and Observed Ability Estimations of Different Starting Rules and Fixed Test Reliability Values in WML and EAP Ability Estimation Methods

		Termination rule: fixed test reliability		
Estimation	Starting rule	SE<0.20	SE<0.30	SE<0.40
WML	-1.50 < b < -0.50 (Easy)	0.981*	0.959*	0.932*
	-0.50 < b < +0.50 (Moderate)	0.982*	0.960*	0.931*
	+0.50 < b < +1.50 (Difficult)	0.979*	0.962*	0.927*
EAP	-1.50 < b < -0.50 (Easy)	0.966*	0.943*	0.894*
	-0.50 < b < +0.50 (Moderate)	0.961*	0.938*	0.887*
	+0.50 < b < +1.50 (Difficult)	0.962*	0.943*	0.899*

**Correlations are significant at 0.01 level.*

As expected, termination rules with smaller standard errors give more comparable results and end up with higher correlation rates but in any occasion the correlation coefficients are above 0.927 for WML and 0.887 for EAP. In all cases, WML ability estimation method produces better expectations to true theta than EAP method.

4.2.2 Determination of Optimum Starting Rule with Fixed Test Length Termination Criteria

Previously, different fixed test reliability values are tested under WML and EAP ability estimations. At the moment again the results of 18 different simulations are indicated in which 3 different fixed test lengths (tests with 15, 25 and 35 items) are tested with 3 different starting rules (start to the test with easy, moderate or difficult item) by 2 different ability estimations: WML and EAP.

The standard error estimations of abilities are the main criteria in fixed length tests. Table 4.5 indicates the statistical information about standard error estimations under WML and EAP methods. Therefore, the comparison of ability estimation methods, starting rules and termination criteria can easily be observed.

Table 4.5 Standard Error Estimations of Monte Carlo Simulations under Different Starting Rules and Fixed Test Length Termination Rules by WML and EAP Ability Estimation Methods

Estimation	Starting rule		Termination rule: fixed test length N=15	N=25	N=35
WML	-1.50 < b < -0.50 (Easy)	Mean	0.357	0.274	0.235
		Min.	0.300	0.238	0.208
		Max.	0.955	0.765	0.372
	-0.50 < b < +0.50 (Moderate)	Mean	0.347	0.272	0.233
		Min.	0.295	0.236	0.207
		Max.	0.929	0.701	0.397
	+0.50 < b < +1.50 (Difficult)	Mean	0.353	0.276	0.236
		Min.	0.303	0.243	0.210
		Max.	0.949	0.644	0.653
EAP	-1.50 < b < -0.50 (Easy)	Mean	0.356	0.281	0.243
		Min.	0.296	0.226	0.172
		Max.	0.713	0.560	0.547
	-0.50 < b < +0.50 (Moderate)	Mean	0.354	0.280	0.242
		Min.	0.290	0.221	0.171
		Max.	0.684	0.631	0.586
	+0.50 < b < +1.50 (Difficult)	Mean	0.358	0.283	0.246
		Min.	0.301	0.229	0.173
		Max.	0.663	0.574	0.523

Mean values of standard error estimations in Table 4.5 point out that WML produce slightly better estimations than EAP. Not surprisingly, more reliable ability estimations are obtained when the number of items in the test is increased. The following table, Table 4.6, shows that how expected scores and observed scores of WML and EAP are correlated to each other.

Table 4.6 Correlation Coefficients between Expected and Observed Scores of Different Starting Rules and Fixed Test Length Values by WML and EAP Ability Estimation Methods

		Termination rule: fixed test length		
Estimation	Starting rule	N=15	N=25	N=35
WML	-1.50 < b < -0.50 (Easy)	0.937*	0.962*	0.973*
	-0.50 < b < +0.50 (Moderate)	0.939*	0.966*	0.971*
	+0.50 < b < +1.50 (Difficult)	0.940*	0.965*	0.975*
EAP	-1.50 < b < -0.50 (Easy)	0.912*	0.949*	0.964*
	-0.50 < b < +0.50 (Moderate)	0.911*	0.943*	0.965*
	+0.50 < b < +1.50 (Difficult)	0.911*	0.948*	0.961*

**Correlations are significant at 0.01 level.*

According to Table 4.6, expected and observed scores of WML and EAP methods are highly correlated to each other. All the correlation coefficients are above 0.90 but WML gives better correlation coefficients with true theta than EAP.

To sum up, results of MC simulations indicate that fixed length test may result with high standard error values which cause low reliability. Even for tests containing 35 items, there exist large SE values such as 0.653 and 0.586 in WML and EAP ability estimations, respectively. Therefore, it will probably be better to fix the test reliability instead of fixing the test length for optimum CAT algorithm. Additionally, among the fixed test reliability termination rules 0.30 give optimum results because 0.20 needs too many items to be administered and 0.40 estimates abilities with rather low precision. Furthermore, defining different starting rules produce almost the same results but during a real CAT administration psychological characteristics of the examinees also need to be considered. Therefore, starting with an easy item will probably be a better choice.

4.2.3 Determination of Item Exposure Rates

In order to determine whether to use an item exposure control strategy in an optimum CAT algorithm, MC simulations are replicated 10 times with 1000 examinees. Each replication is coded as R1, R2, …, R10. The results of the replication study in WML and EAP are given in Table 4.7 and Table 4.8, respectively. These tables contain both the number of items that are not administered to any examinee and the maximum exposure rate of the items.

Table 4.7 Results of 10 Replications in WML to Determine Item Exposure Rates

WML

Criteria	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Items not administered	348	307	291	318	340	325	293	309	338	320
Maximum exposure rate	0.71	0.70	0.71	0.71	0.72	0.70	0.67	0.72	0.70	0.69

Table 4.8 Results of 10 Replications in EAP to Determine Item Exposure Rates

EAP

Criteria	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Items not asked	377	328	259	332	353	334	391	269	365	351
Maximum exposure rate	0.97	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97

Comparison of Table 4.7 and Table 4.8 points out that WML produces lower maximum exposure rates than EAP does. Moreover, almost one third of the items are not given to any examinee in both ability estimations. Therefore, it will be better to use an item exposure control strategy and prefer using WML ability estimation in an optimum CAT algorithm.

4.3 Results of Post-hoc Simulation Studies

In Post-hoc simulations, real responses of the examinees are used to test different scenarios of CAT algorithm. For this purpose, a web-based application is developed by the researcher and coded with PHP programming language over MySQL database. After the program is tested with manual calculations, Post-hoc simulations are implemented on both maximum likelihood (ML) and weighted maximum likelihood (WML) ability estimations. From now on, the results of different simulations will be shared in which ML and WML estimation methods are tested for optimum CAT algorithm.

4.3.1 Determination of Optimum Starting Rule with Fixed Test Reliability Termination Criteria

First Post-hoc simulation is implemented to test ML and WML ability estimation methods under three different starting rules and three different standard error values as a termination rule. As a starting rule, first item is selected either from -1.50 to -0.50 range (easy initial item), -0.50 to 0.50 range (moderate initial item) or 0.50 to 1.50 range (difficult initial item). For the termination, fixed test reliability of SE=0.20, SE=0.30 and SE=0.40 constraints and fixed test length of 15, 25 and 35 items are under control. Table 4.9 contains statistical information of items under fixed test reliability termination rule in ML and WML ability estimations.

Table 4.9 Number of Items Administered under Different Starting Rules and Fixed Test Reliability Termination Rules in Maximum Likelihood (ML) Ability Estimation

Estimation	Starting rule		Termination rule: fixed test reliability		
			SE<0.20	SE<0.30	SE<0.40
ML	-1.50 < b < -0.50 (Easy)	Mean	40.97	16.20	8.78
		Min.	14	7	4
		Max.	163	61	37
	-0.50 < b < +0.50 (Moderate)	Mean	41.52	15.70	8.71
		Min.	13	6	4
		Max.	157	60	30
	+0.50 < b < +1.50 (Difficult)	Mean	41.44	16.12	8.95
		Min.	13	7	5
		Max.	176	69	54
WML	-1.50 < b < -0.50 (Easy)	Mean	41.41	15.85	8.94
		Min.	14	7	4
		Max.	244	65	40
	-0.50 < b < +0.50 (Moderate)	Mean	41.02	15.78	9.01
		Min.	14	7	4
		Max.	172	66	31
	+0.50 < b < +1.50 (Difficult)	Mean	39.93	15.89	8.97
		Min.	40.83	7	5
		Max.	176	76	44

After the abilities of examinees are estimated by ML and WML methods, these ability estimations are correlated with PMS mathematics assessment scores. Table 4.10 figures out the degree of correlations between ML and WML ability estimations and PMS mathematics assessment scores. All the correlations are significant at the 0.01 level.

Table 4.10 Correlations Coefficients of PMS Mathematics Assessment Scores under Different Starting Rules and Fixed Test Reliability Termination Rules

		Termination rule: fixed test reliability		
Estimation	Starting rule	SE<0.20	SE<0.30	SE<0.40
ML	-1.50 < b < -0.50 (Easy)	0.776*	0.721*	0.668*
	-0.50 < b < +0.50 (Moderate)	0.776*	0.732*	0.670*
	+0.50 < b < +1.50 (Difficult)	0.780*	0.716*	0.701*
WML	-1.50 < b < -0.50 (Easy)	0.794*	0.734*	0.689*
	-0.50 < b < +0.50 (Moderate)	0.801*	0.753*	0.690*
	+0.50 < b < +1.50 (Difficult)	0.792*	0.741*	0.673*

**All correlations are significant at the 0.01 level.*

4.3.2 Determination of Optimum Starting Rule with Fixed Test Length Termination Criteria

Second Post-hoc simulation is designed to test the effect of fixed length termination rule under 15, 25 and 35 items. Since test length is fixed, standard error values are varied in ability estimations. Table 4.11 gives the detailed analysis of standard error values of ML and WML ability estimations.

Table 4.11 Standard Error Estimations of Post-hoc simulations under Different Starting Rules and Fixed Test Length Termination Rules in ML Ability Estimation

Estimation	Starting rule		Termination rule: fixed test length		
			N=15	N=25	N=35
ML	-1.50 < b < -0.50 (Easy)	Mean	0.289	0.240	0.217
		Min.	0.192	0.160	0.147
		Max.	0.686	0.675	0.342
	-0.50 < b < +0.50 (Moderate)	Mean	0.301	0.245	0.213
		Min.	0.205	0.161	0.138
		Max.	0.564	0.487	0.399
	+0.50 < b < +1.50 (Difficult)	Mean	0.298	0.244	0.218
		Min.	0.207	0.160	0.143
		Max.	0.614	0.418	0.417
WML	-1.50 < b < -0.50 (Easy)	Mean	0.289	0.236	0.211
		Min.	0.189	0.152	0.138
		Max.	0.738	0.554	0.394
	-0.50 < b < +0.50 (Moderate)	Mean	0.288	0.236	0.212
		Min.	0.190	0.158	0.138
		Max.	0.714	0.565	0.496
	+0.50 < b < +1.50 (Difficult)	Mean	0.288	0.237	0.212
		Min.	0.192	0.149	0.136
		Max.	0.756	0.554	0.434

The table given below shows how PMS mathematics assessment scores and Post-hoc simulation ability estimations obtained from different fixed-test lengths are correlated. All the correlations are significant at the 0.01 level.

Table 4.12 Correlations Coefficients of PMS Mathematics Assessment Scores under Different Starting Rules and Fixed Test Length Termination Rules

Estimation	Starting rule	Termination rule: fixed test length		
		15 items	25 items	35 items
ML	-1.50 < b < -0.50 (Easy)	0.748*	0.801*	0.822*
	-0.50 < b < +0.50 (Moderate)	0.765*	0.801*	0.816
	+0.50 < b < +1.50 (Difficult)	0.758*	0.803*	0.827
WML	-1.50 < b < -0.50 (Easy)	0.763*	0.812*	0.832*
	-0.50 < b < +0.50 (Moderate)	0.766*	0.814*	0.833*
	+0.50 < b < +1.50 (Difficult)	0.773*	0.811*	0.830*

**All correlations are significant at the 0.01 level.*

As stated before, standard error values are directly related to the reliability of the test scores. In fixed-length testing, standard error of a score varies for ability estimations but the examiner is sure that same number of items is taken for all examinees. On the other hand, fixation of standard error value guarantees the reliability of ability estimations in variable length tests but the number of items administered varies for each examinee. Additionally, the analysis of Table 4.10 and Table 4.12 state that different starting rules provide parallel correlation coefficients which leads to the fact that the simulation is invariant of the difficulty of initial item.

4.3.3 Determination of a Need to Content and Item Exposure Control Strategies

PMS item bank consists of items from 4 different areas: geometry (GE), measurement (ME), numbers (NU) and probability & statistics (PS). In a PMS administration, examinees take almost the same number of items but the number of items in these areas may differ in each grade level. For example, geometry testlet with 18 items, measurement testlet with 20 items, numbers testlet with 27 items and probability & statistics testlet with 13 items constitutes 78 items of a 7^{th} grade PMS mathematics assessment. As a result of previously designed linear testlets, content and exposure rates of items are under control in PMS. On the other hand, it is not possible to check these constraints in a CAT administration without using any content control or item exposure control strategy. In order to determine the need to item exposure and content controls, third simulation study is designed. For this purpose, optimum CAT algorithm obtained from prior simulation studies is set. Exposure rates and content of items given in optimum CAT simulation are shown in Figure 5 and Table 4.13, respectively.

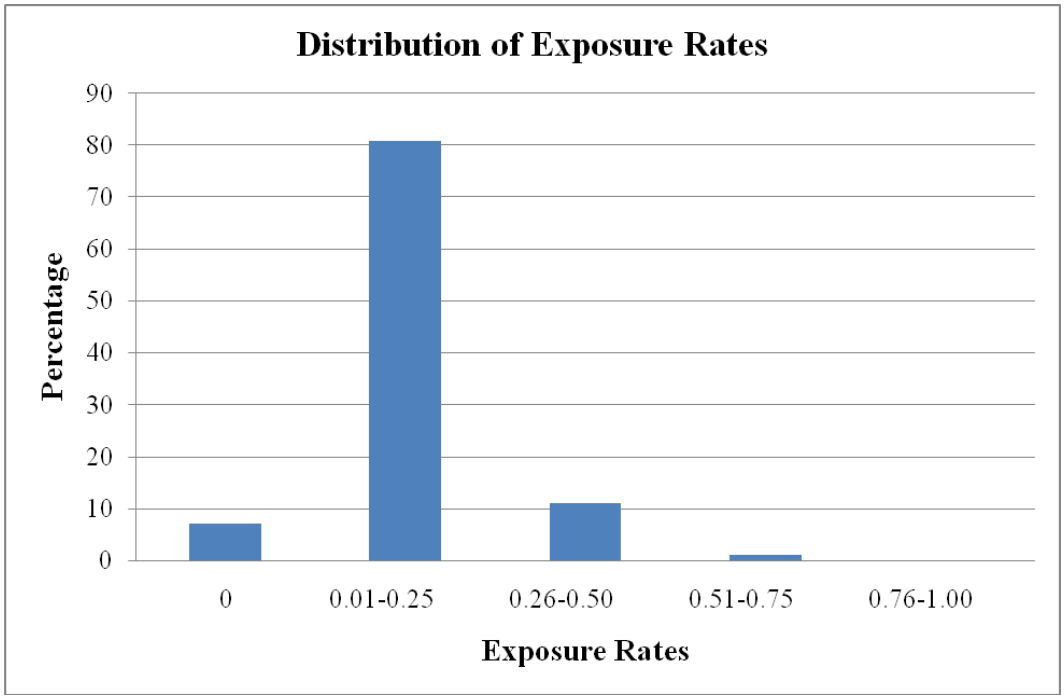

Figure 5. Distribution of Exposure Rates in CAT Simulation

Table 4.13 Content Analysis of Items in CAT Simulation

	GE	ME	NU	PS
Content distribution of item bank (%)	24.3%	24.5%	40.0%	11.2%
Content distribution of simulations (%)	18.7%	30.3%	42.8%	8.2%
Percentage of administrations not implemented	12.0%	5.1%	0.4%	43.8%

The exposure rates given in Figure 5 indicate that there is a need to item exposure control strategy because more than 10% of the items are over exposed and almost

7% of the items are under exposed. Moreover, Table 4.13 shows that without a content control 12.0% of the pupils do not face with any Geometry item, 5.1% with any Measurement item, 0.4% with any Numbers item and 43.8% with any Probability and Statistics item. This result provide information about the use of content control for optimum CAT algorithm. Therefore, the effect of using content and exposure control strategies on estimating the ability estimations is tested as third Post-hoc simulation. In these Post-hoc simulations 4 different situations are tested: (1) no use of content and exposure control, (2) use of only content control, (3) use of only exposure control and finally (4) using both content and exposure controls are simulated. The number of items administered and correlation coefficients between ability estimations of two occasions, both in PMS administrations and Post-hoc simulations, are denoted in Table 4.14.

Table 4.14 Correlations of Ability Estimations between PMS Assessment and Post-Hoc CAT Simulation under the Use of Content and Exposure Control

	Average Number of Items Administered	Correlation with Weighted Mean
No content and exposure control	15.78	0.753*
Content control only	12.77	0.730*
Exposure control only	20.74	0.777*
Both content and exposure control	17.09	0.771*

In Figure 6, the effect of using item exposure control strategy on exposure rates can easily be observed. Small red triangles represent the previous exposure rates and blue squares symbolize exposure rates after the control strategy is implemented. Post-exposure rates are more compact and closer to 0. However, the previous exposure rates are either distributed to upper parts of the scale or scattered on horizontal axis. Obviously, an exposure rate of 0 (zero) means that the item is not implemented in any testing occasion.

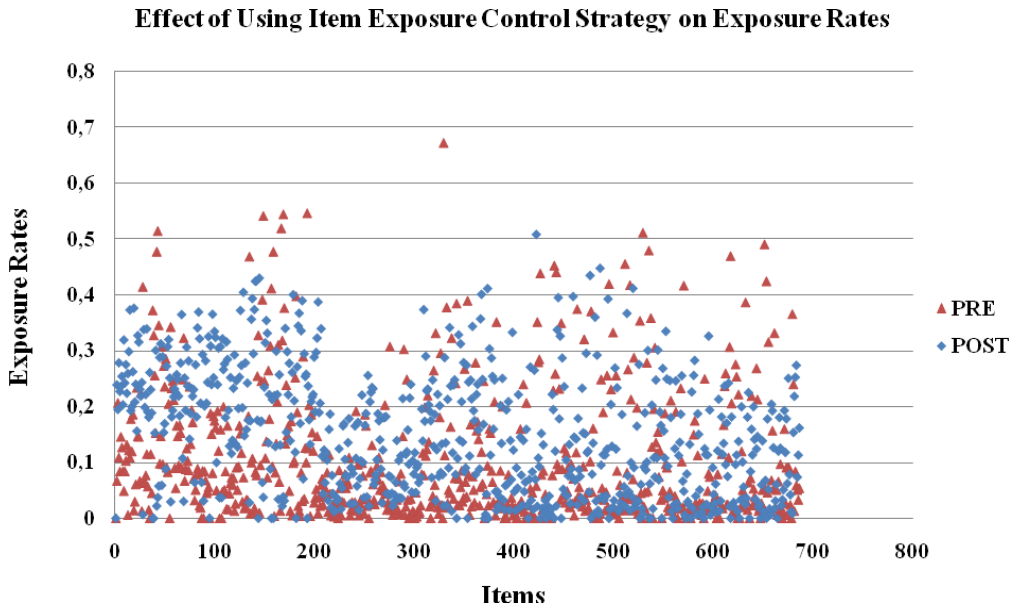

Figure 6. Effect of Item Exposure Rates Before and After Exposure Control

Figure 6 indicates the exposure rates of items. The exposure rates labeled as PRE are obtained from the simulations in which no item exposure control strategy is used whereas POST exposure rates are obtained from the simulation in which

Sympson & Hetter (SH) is set as item exposure control strategy. As shown in the figure, using SH exposure control strategy decrases the number of items which are not administered in any occasions and also centralizes the outliers of the item exposure rates.

According to the results of previous Post-hoc simulations, an optimum CAT algorithm for PMS mathematics assessments uses WML ability estimation, implements easy initial item, terminates after at least 0.91 fixed test reliability is satisfied (when SE<0.30), uses both item exposure and content control strategies. Therefore, the final Post-hoc simulation is designed so that a fixed number of open ended-items are administered at the end of the test. The reason for simulating such a procedure is to create the most compatible real life administration of CAT. Correlation coefficients are given in Table 4.15.

Table 4.15 Correlation Coefficients of Ability Estimations between PMS Mathematics Assessments and Optimum CAT Algorithm Simulation Results

	GE	ME	NU	PS	Weighted Mean
Optimum CAT Algorithm	0.697*	0.722*	0.755*	0.641*	0.840*

**All correlations are significant at the 0.01 level.*

As shown in Table 4.15, implementing most informative 5 open ended items after fixed test reliability of 0.91 (SE<0.30) satisfied by multiple choice items provides high correlation of 0.840 with PMS scores.

Additionally, ability estimations of Post-hoc simulations can be used to determine pupils' proficiency levels with respect to the cut-off points defined PMS mathematics assessment. Afterwards, the percentages of correct proficiency levels in all sub-domains are evaluated and given in Table 4.16.

Table 4.16 Percentages of Correct Proficiency Levels with Different Termination Criteria in Post-hoc Simulations

	GE	ME	NU	PS
Correct Proficiency Level with SE<0.30	44.5%	56.9%	47.8%	61.6%
Correct Proficiency Level with SE<0.20	45.4%	64.0%	55.2%	66.1%

Table 4.16 shows that Probability and Statistics (PS) has the highest correct percentages under two different test termination criteria. For instance, 56.9% to 61.6% ME and PS proficiency levels can be estimated correctly by the optimum CAT algorithm developed by the researcher. However, correct percentages of proficiency levels for GE and NU are around 44.5% and 47.8%. When more reliable termination criteria is defined as 0.96 (SE<0.20), there exist an increase in the percentages of correctly determined proficiency levels.

CHAPTER 5

CONCLUSION AND DISCUSSION

This chapter mainly summarizes the results of the current study and discusses the findings. As it is mentioned in the first chapter, purpose of the present study is to investigate the applicability of CAT to PMS mathematics assessments. Therefore, a live CAT administration is designed and implemented to 454 students by using 507 multiple choice items. Afterwards, the researcher expands the item bank with open ended items and conducts several Monte Carlo and Post-hoc simulations to determine the optimum CAT algorithm. In Monte Carlo simulations, 1,000 examinees are taken in each run. Additionally, 1,000 examinees are randomly selected from the data of 12,156 pupils in Post-hoc simulation phase and also this sample is fixed eliminate the other factors.

5.1 Summary of Findings

Live CAT Administration:

- Correlation coefficients of ability estimations between PMS mathematics assessments (GE, ME, NU and PS) and live CAT administration are between 0.65 and 0.70 range. However, live CAT ability estimations show higher correlation with weighted mean of PMS mathematics sub-domain scores which is 0.75.

- In live CAT administration, fixed test reliability termination rule with SE<0.30 is used. Average number of items used in the study is 19.
- Item exposure control is not used in live CAT administration phase. Outcomes of live CAT administration indicate that 18 items (4%) have an exposure rate bigger than 0.30 and 283 items (55.8%) do not exist in any administration.
- Content control is also not used in live CAT administration phase. Almost 30% of the examinees do not face with even one probability and statistics (PS) item in their computerized adaptive test sessions.
- 31.7% to 64.8% of the pupils' GE, ME, NU and PS proficiency levels can be estimated correctly in live CAT administration.

Monte Carlo (MC) Simulations:

- MC simulations indicate that change in the difficulty of the initial item does not provide different ability estimations.
- WML ability estimation method requires fewer items than EAP to estimate the ability values with the same precision.
- WML ability estimations provide slightly higher correlations with PMS mathematics assessment scores than EAP.
- An increase in the number of administered items ends up with higher precise of measurement.

Post-hoc Simulations:

- Variations on initial item difficulties provide similar test lengths and ability estimations.
- For fixed test reliability termination rule, approximately 41 items are needed to satisfy ability estimations with 0.20 SE, 16 items with 0.30 SE and 9 items with 0.40 SE.
- For fixed test length termination rule, 15 items estimate the abilities with an average of 0.30 SE, 25 items with 0.24 SE and 35 items with 0.21 SE.
- WML provides better ability estimations than ML.
- Using content control in CAT algorithm produces a decrease in the average number of items used in CAT administrations.
- Using item exposure control substantially increases the average number of items used in CAT administrations.
- Ability estimations obtained from CAT algorithm that uses both item exposure and content controls gives high correlations around 0.77.
- Implementing open-ended items after multiple choice items provide a correlation coefficient of 0.84 with PMS mathematics assessment.
- 44.5% to 61.6% of the proficiency levels in mathematics subdomains can be estimated correctly in Post-hoc simulations by the optimum CAT algorithm.

5.2 Live CAT Administration Phase

After 507 multiple choice items are calibrated by BILOG-MG, a fixed CAT algorithm is designed which uses ML ability estimation method and maximum information item selection strategy. Computerized adaptive testing session of each individual starts with a random item with moderate difficulty (having item difficulty values between -0.1 and 0.1) and terminates after standard error is lower than 0.30. No item exposure and content control is used within the algorithm. 454

students take live CAT administration from 7 different schools (4 of them are private schools and 3 of them are state schools).

Findings of this phase are in parallel to the literature. For instance, correlation coefficients between the linear and CAT versions of a test are between 0.74 and 0.94 in the literature (Kalender, 2011; Molina, 2009; Gafni, Cohen, Roded, Baumer & Moshinsky; 2009; Iseri, 2002; Kingsbury, 2002). Although, correlation coefficients of ability estimations between live CAT administration and PMS mathematics assessment lie between 0.65 and 0.70, the correlation between the weighted mean score of sub-domains and live CAT ability estimations is found to be 0.75. Furthermore, a CAT study conducted by Eggen & Straetmans (2000) indicates that 52.8% of the items are never selected without using item exposure controls in a CAT administration. Eggen (2001) provide similar results that 66% of the items are not asked to any of the examinee. Similar finding exist in this phase such that 55.8% (283 of 507) items do not exist in the CAT sessions of examinees.

5.3 Monte Carlo Simulations Phase

As indicated previously, items that are used in Monte Carlo (MC) simulations are both multiple-choice and open-ended type. Item calibrations are conducted by MIRT program developed by Glas. 1,000 examinees are randomly selected in each run of Monte Carlo simulations. Instead of designing empirical live CAT studies, most of the researchers in the literature prefer using MC simulations both to obtain information about the characteristics of item bank and to determine optimum CAT algorithm for the data because it is easier to use. However, MC does not allow the comparison of PMS mathematics assessment scores with ability estimations obtained from simulations because it is impossible to match an ability estimation of a MC simulation with Turkish PMS mathematics assessment score. Therefore, instead of comparing ability estimations, discussion is mainly

focused on the optimum ability estimation methods, starting rules and termination criteria.

Starting rule is simply related to the selection of initial item in a CAT algorithm. Since the algorithm does not have any information about an examinee at first, logically initial item needs to have an average difficulty. However, testing programs generally want initial successful experience of test takers so they prefer to administer a slightly easier first item than average (Mills & Stocking, 1996). Results of MC simulations show that different initial item difficulties of CAT algorithms produce similar ability estimations.

Termination criterion of a CAT algorithm determines when to stop implementing items. In MC simulations, ability estimations obtained from different fixed length tests containing 15, 25 and 35 items and different fixed test reliability values of 0.84 (for SE=0.40), 0.91 (for SE=0.30) and 0.96 (for SE=0.20) are compared. A general fact in the literature states that an increase in the number of items in a test ends up with higher reliability. Findings of MC simulations regarding termination criteria are not different. In order to satisfy 0.20 standard error, approximately 50 to 72 items need to be implemented. 20 to 22 items are needed for 0.30 standard error and 11 to 12 items are needed for 0.40 standard error. In other words, a decrease in the value of standard error ends up with an increase in the number of items administered so more precise ability estimations are obtained.

In addition, WML and EAP methods are simulated to determine the optimum ability estimation method in MC simulations. Findings are similar to the studies in the literature. For example, Wang and Wang (2001) conduct a Monte Carlo study to compare ML, WML, EAP an MAP ability estimation methods. There are no significant differences in estimations among different ability estimation methods but WML gives slightly better results. The findings of MC simulations about ability estimations indicate that WML provides slightly higher correlations with true scores than EAP does.

Finally, replication studies aiming to determine the item exposure rates in MC simulations declares that 30%-40% of the items are never used. Maximum exposure rates are around 0.70 for WML and around 0.97 for EAP. These findings indicate that an item exposure control strategy is needed for optimum CAT algorithm.

5.4 Post-hoc Simulations Phase

Rather than MC, Post-hoc simulations use real data and provide comparison of ability estimation outcomes with Turkish PMS mathematics assessment scores. Each pupil's ability estimation is matched with mathematics scores obtained from PMS and then correlation coefficients are used to determine the optimum CAT algorithm. Before conducting simulations, a sample of 1,000 examinees are randomly selected fixed to observe the correlation without any extra factors.

However, like MC simulations, Post-hoc simulations test the efficiency of three different initial item difficulty values and obtain parallel outcomes with MC such that different initial item difficulty values generates similar ability estimations.

CAT algorithms have either fixed test length termination or fixed test reliability termination criteria. To determine the optimum termination rule, each type of termination is tested under 3 different alternatives. That is to say, fixed length test termination rule is compared with 15, 25 and 35 items whereas fixed test reliability is compared with 0.20 (0.96 reliability), 0.30 (0.91 reliability) and 0.40 (0.84 reliability) standard error values. Moreover, ability estimations obtained from six different termination criteria are compared with PMS mathematics assessment scores. According to the results of simulations aiming to determine the optimum test termination, fixed test reliability termination rule with 0.30 standard error ends up with comparable ability estimations by using approximately 16 items. Kalender (2011) uses fixed test reliability termination rule with a SE value of 0.30 whereas Molina (2009) and Iseri (2002) uses 0.32 SE. Indeed, 0.20

standard error provides higher correlations but the average number of items used in the test increases from 16 to 41.

On the other hand, fixed length tests estimates ability with different precise of measurement such that by using 15 items each ability is estimated with 0.30 SE, by 25 items 0.24 SE and by 35 items 0.21 SE. Some other researchers use fixed test length termination criteria of 10 to 30 items (Costa, Karino, Moura & Andrade, 2009; Dodd, Koch & De Ayala, 1989; Stocking, 1987).

Eggen (2001) states that interferences to adaptive algorithm affect the efficiency of CAT but the results of the study about using item exposure control of WISCAT test package conducted in Cito indicates that using item exposure has a positive impact on item bank. Additionally, the comparative study developed by Eggen & Straetmans (2000) state that CAT algorithms in which item exposure and content control strategies are used provide better ability estimations. Result of Post-hoc simulations are also in parallel to the findings within the literature (Eggen & Straetmans, 2000; Costa, Karino, Moura & Andrade; 2009). Using only item exposure control increases the average number of items sharply but Sympson&Hetter (SH) item exposure control strategy with Kingsbury&Zara content control provides comparable ability estimations with PMS mathematics assessment scores.

The last step of Post-hoc simulations is related to the design of computerized adaptive test. A simulation which is familiar to a live CAT administration is designed by using the optimum CAT algorithm. Multiple choice items are given until standard error of ability estimation is below 0.30. Next, the most informative 5 open-ended items are given to each individual. That is to say, after abilities are estimated by multiple-choice items with 0.30 standard error, 5 most informative open-ended items are identified regarding content. The correlation coefficient of ability estimations with PMS weighted mean score is 0.84.

This study also focuses on the correct distribution of ability estimations into proficiency levels. 44.5% to 61.6% of the ability estimations are correctly located into the proficiency levels by using the optimum CAT algorithm in Post-hoc simulations. In a similar study, Verschoor and Straetmans (2010) investigates the use of computerized adaptive placement test instead of administering a two stage testing in a mathematics courses in adult basic education. However, the results of this study indicate that 76.0% to 91.1% of the proficiency levels in are estimated correctly. Probably, the range between each cut-off score has an impact on the percentages of correctly determined proficiency levels.

5.5 Applicability of CAT in Turkish PMS

This initial study conducted for the mathematics assessment of PMS produces promising results. Before conducting the current study, PMS assessments have already been based on IRT framework and have been using computer based tests. Therefore, such reasons foster the use of CAT in Turkish PMS.

Using CAT in PMS is also recommended in the literature. According to Glas & Geerlings (2009) state that is advantageous to use CAT in PMS because of two main reasons: measurement efficiency and possibility of testing on demand. Since the present study gives parallel findings to the literature and there exist studies recommending the use of CAT in PMS, computerized adaptive testing can be applied to Turkish PMS mathematics assessment. According to the findings of simulation studies, CAT ability estimations give higher correlation with PMS mathematics assessment scores if the CAT algorithm has the following properties:

- Although different difficulty values of the first item give similar results, it is better to start with an easy item in order to lower test anxiety.
- Rather than fixed test lengths, fixed test reliability with SE<0.30 provides optimum termination rule with an average of approximately 16 items.

- WML ends up with better ability estimations than ML and EAP. Hence, using WML ability estimation method in CAT algorithm provides more promising results.
- Using item exposure and content control has a positive impact on ability estimations and validity issues.
- For a practical CAT administration, a set of multiple choice items needs to be given until a termination criterion is satisfied. Afterwards a fixed set of open-ended items from each sub-domains are administered. In the present study, multiple-choice items are followed by 5 open-ended items. This number can be increased according to the purpose of the assessment.

5.6 Limitations of the Study

In the study, group specific item parameters are not considered, i.e. the existence of Differential Item Functioning (DIF) items is not taken into account. Determination of DIF items in PMS mathematics assessments can be studied.

In Post-hoc simulations phase, ability estimations of some examinees can be evaluated after a large number of items are implemented. Especially for high and low ability students the risk of taking too many items exists. In order to overcome such limitation of CAT administration, PMS item bank can be expanded with the items having high and low difficulty values.

Final limitation of the current study is related to the ability estimations in each sub-domain of mathematics. As it is mentioned earlier, PMS mathematics assessment is composed of 4 sub-domains: geometry (GE), measurement (ME), numbers (NU) and probability and statistics (PS). At the end of a PMS administration, each examinee has 4 scores for each sub-domain. On the other hand, CAT administration provides unique ability estimation since items from each sub-domain are calibrated together to produce a large item bank. Ability

estimations provide high correlations with weighted mean of each sub-domain but comparatively lower correlations with each PMS sub-domain scores. Increasing the number of items of each sub-domain and designing CAT studies focusing on ability estimations calculated for each sub-domain can give more comparable results with PMS mathematics assessment.

5.7 Suggestions for Further Research

First of all, the present study is focused on PMS mathematics assessments of 4^{th}, 5^{th} and 6^{th} grades. Different CAT studies using items from different content areas such as Turkish literature, science and technology and social sciences can be designed on the optimum algorithm obtained from this study.

In the present study, CAT algorithm does not have any information about an individual at first so initial ability estimation of each examinee is taken as zero. On the other hand, if pupils' previous ability estimations are entered to CAT program, algorithm of the program probably provides ability estimations with fewer items. As a result, a research design dealing with different initial ability estimations can be conducted for further research.

The effectiveness of different ability estimation methods such as ML, WML and EAP is compared within different scenarios in the current study but not MAP. Although most studies state that there is no significant difference between the ability estimations obtained from ML, WML, EAP and MAP, a study investigating the effect of MAP on Turkish PMS mathematics assessment can be evaluated.

Additionally, the present study focuses only on maximum information item selection method. CAT algorithms using different item selection methods such as Kullback-Leibler information (KL) or maximum expected information (MEI) can be tested.

Finally, some open-ended mathematics items in PMS assessments are graded by partial credits. Therefore, a study using logistic models for multiple-choice items and partial credit model (PCM) for open-ended items can be designed to produce comparable ability estimations.

REFERENCES

Babcock, B. & Weiss, D. (2009). Termination criteria in computerized adaptive tests: variable-length cats are not biased. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Birnbaum, A. (1968). Some Latent Trait Models. In F.M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Burket, G. R. (1996). PARDUX (Version 4.1). Monterey, CA:CTB/McGraw-Hill.

Camilli, G. (1994). Origin of the Scaling Constant D=1.7 in Item Response Theory. Journal of Educational and Behavioral Statistics, Vol. 19, No. 3, pp. 293-295.

Costa, D. R., Karino, C. A., Moura, F. A. S., Andrade, D. F. (2009). A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved 05.03.2011 from http://www.psych.umn.edu/psylabs/CATCentral/

DeMars, C. (2010). Item Response Theory. Understanding Statistics, Measurement. Oxford; New York: Oxford University Press.

Dodd, B. G., Koch W. R. & De Ayala, R. J. (1989). Operational Characteristics of Adaptive Testing Procedures using the Graded Response Model. Applied Psychological Measurement, 13, 129-143.

Eggen, T. J. H. M. (2007). Choices in CAT Models in the Context of Educational Testing. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved 25.04.2012 from http://www.psych.umn.edu/psylabs/CATCentral/

Eggen, T. J. H. M. (2004). Contributions to the Theory and Practice of Computerized Adaptive Testing. Dissertation. Print Partners Ipskamp B.V., Enschede.

Eggen, T. J. H. M. (2001). Overexposure and Underexposure of Items in Computerized Adaptive Testing. Measurement and Research Department Reports, 2001-1. Citogroep, Arnhem.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized Adaptive Testing for Classifying Examinees into Three Categories. Educational and Psychological Measurement, Vol. 60, pp. 713-734.

Eggen, T. J. H. M., & Verhelst, N. D. (2011). Item calibration in Incomplete Designs. Psicológica, Vol. 32, pp. 107-132.

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal Testing with Easy or Difficult Items in Computerized Adaptive Testing. Applied Psychological Measurement, 2006, Vol. 30 (5), p. 379-393.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., Link, V. (1998). Calibration and Scoring of Tests with Multiple-Choice and Constructed Item Types. Journal of Educational Measurement. Vol. 35, No. 2, pp 137-154.

ETS (1994). Computer-Based Tests: Can They be Fair to Everyone? Princeton, NJ: Educational Testing Service.

Georgiadou E., Triantafillou E., Economides A. A. (2007). A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. The Journal of Technology, Learning and Assessment. Vol. 5, No. 8. Retrieved 09.02.2012 from http://www.jtla.org.

Glas, C. A.W., Geerlings, H. (2009). Psychometric Aspects of Pupil Monitoring Systems. Studies in Educational Evaluation 35, pp. 83-88.

Glas, C. A. W. (2010). Preliminary Manual of Software Program: Multidimensional Item Response Theory (MIRT). Retrieved March 28, 2012, from University of Twente, Department of Research Methodology, Measurement and Data Analysis Website http://www.utwente.nl/gw/omd/afdeling/temp_test/mirt-manual.pdf

Gu, L., Reckase M. D. (2007). Designing Optimal Item Pools for Computerized Adaptive Tests with Sympson-Hetter Exposure Control. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved 05.04.2012 from http://www.psych.umn.edu/psylabs/CATCentral/

Haley, D. C. (1952). Estimation of the Dosage Mortality Relationship when the Dose is Subject to Error. Technical Report No. 15. Stanford. Calif.: Stanford University. Applied Mathematics and Statistics Laboratory.

Hambleton, R.K., Swaminathan H. & Rogers H. J. (1991). Fundamentals of Item Response Theory. Measurement Methods for the Social Sciences Series. Sage Publications, Inc.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement: Issues and Practice, Vol. 12, Issue 3, pp. 38-47.

Hoijtink, H. & Boomsma, A. (1995). On Person Parameter Estimation in the Dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch Models: Foundations, Recent Developments and Applications (pp. 53-68). New York: Springer-Verlag.

Hoijtink, H., Vollema, M. (2003). Contemporary Extensions of the Rasch Model. Quality & Quality. Vol. 37, pp 263-276. Kluwer Academic Publishers.

Imai, S., Ito, S., Nakamura, Y., Kikuchi, K., Akagi, Y., Nakasono, H., Honda, A., & Hiramura, T. (2009). Features of J-CAT (Japanese Computerized Adaptive Test). In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved 03.05.2012 from http://www.psych.umn.edu/psylabs/CATCentral/

Is Guzel, C., Berberoglu, G., Demirtasli, N., Arikan, S., Ozgen Tuncer, C. (2009). Ogretim Programlarinin Ogrenme Ciktilari Acisindan Degerlendirilmesi, Cito Egitim: Kuram ve Uygulama, Sayi 6, pp 9-30.

Iseri, A. I. (2002). Assessment of Students' Mathematics Achievement Through Computer Adaptive Testing Procedures. Unpublished doctoral dissertation. Middle East Technical University, Turkey.

Kalender, I. (2011). Effects of Different Computerized Adaptive Testing Strategies on Recovery of Ability. Unpublished PhD Thesis, Middle East Technical University, Turkey.

Kaptan, F. (1993). Yetenek Kestiriminde Adaptive (Bireyselleştirilmiş) Test Uygulaması ile Geleneksel Kağıt-Kalem Testi Uygulamasının Karşılaştırılması. Unpublished PhD Thesis, Hacettepe University, Turkey.

Karantonis A., Sireci S. G. (2006). The Bookmark Standard-Setting Method: Literature Review. Educational Measurement: Issues and Practice. Vol. 25, Issue 1, pp 4-12.

Kingsbury, G. G. & Hauser, C. (2004). Computer Adaptive Testing and No Child Left Behind. A Paper For the Session: Exploration of Pertinent Assessment Issues in Large Scale Testing Environments in the 2004 Annual Meeting of the American Educational Research Association, San Diego, CA.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. Applied Measurement in Education. Vol. 2, No. 4, pp 359-375.

Kingsbury, G. G. & Zara, A. R. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests. Applied Measurement in Education, 4, 241-261.

Kingsbury, G. G. (2002). An Empirical Comparison of Achievement Level Estimates from Adaptive Tests and Paper-and-Pencil Tests. Paper presentation at the Annual Meeting of the American Educational Research Association, New Orleans, LA. Retrieved May 3, 2012 from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ki02-01.pdf

Koklu, N. (1990). Klasik Test Teorisinie Göre Geliştirilen Tailored Test ile Grup Testi Arasında bir Karşılaştırma. Unpublished PhD Thesis. Hacettepe University, Turkey.

Kolen, M. J., Brennan R. L. (1995). Test Equating Methods and Practices. Springer Series in Statistics, Springer-Verlog New York, Inc.

Lilley M. & Barker, T. (2004). A Computer-Adaptive Test That Facilitates the Modification of Previously Entered Responses: An Empirical Study. Proceedings of the 2004 Intelligent Tutoring Systems Conference, Lecture Notes in Computer Science 3220, pp. 22-33.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associate, Inc.

Meijer, R. R., Nering M. L., (1999). Computerized Adaptive Testing: Overview and Introduction. Applied Psychological Measurement, Vol. 23, No. 3, pp. 187-194, Sage Publications, Inc.

Mills, C. N. & Stocking, M. L. (1996). Practical Issues in Large-Scale Computerized Adaptive Testing. Applied Measurement in Education, Vol. 9 (4), pp. 287-304.

Moelands, H. (2010). Computerized Adaptive Testing in the Monitoring and Evaluation System for Primary Education in the Netherlands. 36th International Association of Educational Assessment Conference Paper.. Retrieved from: http://www.iaea.info/papers.aspx?id=78.

Molina, M. T. (2009). A Computer-Adaptive Vocabulary Test. Indian Journal of Applied Linguistics, Vol. 35, No. 1, pp. 121-138. Bahri Publications.

Muraki E. & Bock R. D. (1993). PARSCALE: IRT Based Test Scoring and Item Analysis. Chicago, IL: Scientific Software International.

Onder, I. (2007). An Investigation of Goodness of Model Data Fit. Hacettepe Universitesi Eğitim Fakültesi Dergisi, 32, pp 210-220.

Owen, R. J. (1969). A Bayesian Approach to Tailored Testing (RB-69-92). Princeton, NJ: Educational Testing Service.

Ozgen Tuncer, C. (2008). Cito Turkiye Ogrenci Izleme Sistemi (OIS) ve OIS'de Soru Gelistirme Sureci. Cito Egitim: Kuram ve Uygulama, Tanitim Sayisi, pp 22-26.

Powers, D. E. (2001). Test Anxiety and Test Performance: Comparing Paper-Based and Computer-Adaptive Versions of the Graduate Record Examinations (GRE) General Test. Journal of Educational Computing Research, Vol. 24, No. 3, pp. 249-273. Baywood Publishing Co., Inc.

Rasch, G. (1960). Probabilistic Models for some Intelligence and Attainment Tests. Copanhagen: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multi-Factor Tests: Results and Implications. Journal of Educational Statistics, 4, 207-230.

Revuelta, J. & Ponsada, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. Journal of Educational Measurement, 38, 311-327.

Rudner, L. M. (2010). Implementing the Graduate Management Admission Test Computerized Adaptive Test. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing, (pp 151-165). Statistics for Social and Behavioral Sciences. Springer.

Scheerens, J., Glas, C., & Thomas, S. (2003). Educational Evaluation, Assessment, and Monitoring: A Systemic Approach. Lisse: Swets & Zeitlinger.

Schnipke, D. L., Green, B. F. (1995). A Comparison of Item Selection Routines in Linear and Adaptive Tests. Journal of Educational Measurement, Vol. 32, No.3, pp. 227-242.

Schultz, K., and Whitney, D. (2005). *Measurement theory in action.* Thousand Oaks, CA: Sage Publications.

Shermis, M. D., Fulkerson, J., & Banta, T. W. (1996). Computerized Adaptive Math Tests for Elementary Talent Development Selection. Roeper Review, Vol. 19, Issue 2, pp. 91-95.

Stocking, M. L. (1987). Two Simulated Feasibility Studies in Computerized Adaptive Testing. Applied Psychology: An International Review, Vol. 36, pp. 263-277.

Tang, K. L. & Eignor D. R. (1997). Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL Items Using IRT Models. TOEFL Technical Report TR-13. Educational Testing Service, Princeton, New Jersey.

Van der Linden, W. J. (1995). Advances in Computer Applications. In T. Oakland & R. K. Hambleton (Eds.), International Perspectives on Academic Assessment, (pp. 105-124). Kluwer Academic Publishers.

Van der Linden, W. J. (2001). Computerized Test Construction. Research Report. Twente University, Enschede (Netherlands). Faculty of Educational Science and Technology.

Van der Linden, W. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing, (pp 3-30). Statistics for Social and Behavioral Sciences. Springer.

Veerkamp, W. J. J. (1996). Statistical Methods for Computerized Adaptive Testing. Phd Thesis, University of Twente.

Verhelst, N.D. and Glas, C.A.W. (1995). The One Parameter Logistic Model. In G.H. Fischer and I.W. Molenaar (Eds.), Rasch Models: Foundations, Recent Developments and Applications. pp. 215-238. New York: Springer Verlag.

Verschoor, A. J. & Straetmans, G. J. J. (2010). MATHCAT: A Flexible Testing System in Mathematics Education for Adults. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing, (pp 137-149). Statistics for Social and Behavioral Sciences. Springer.

Veldkamp, B. P., Verschoor, A. J. & Eggen, T. J. H. M. (2007). A Multiple Objective Test Assembly Approach for Exposure Control Problems in Computerized Adaptive Testing. Measurement and Research Department Reports, 2007-1. Cito Arnhem.

Vlug, K. M. F. (1997). Because Every Pupil Counts: The Success of the Pupil Monitoring System in the Netherlands. Education and Information Technologies, Vol. 2, Issue 4, pp. 287-306.

Wainer, H. (2000). Computerized Adaptive Testing: A Primer. Mahvah, NJ: Erlbaum.

Wainer, H., Dorans, N. J., Green B. F., Steinberg, L., Flaugher R., Mislevy, R. J. & Thissen, D. (1990). Computerized Adaptive Testing : A Primer. Lawrence Erlbaum Associates, Publishers.

Wang, S. & Wang, T. (2001). Precision of Warm's Weighted Likelihood Estimates for a Polytomous Model in Computerized Adaptive Testing. Applied Psychological Measurement, 25(4), 317–331.

Warm T.A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika, 54, 427-450.

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. Applied Psychological Measurement, 6, 473-492.

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.

Wingersky, M. S. (1983). LOGIST: A Program for Computing Maximum Likelihood Procedures for Logistic Test Models. In R. K. Hambleton (Ed.) Applications of Item Response Theory. Vancouver, B. C.: Educational Research Institute of British Columbia.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). BILOG-MG [Computer Software]. Chicago: Scientific Software International.

APPENDIX A

ITEM PARAMETERS CALIBRATED BY BILOG-MG, OPLM AND MIRT

Table A.1 provides item discrimination and item difficulty parameters calibrated by BILOG-MG, OPLM and MIRT. OPLM_A and OPLM_B indicate the item parameters which are used in PMS administrations. OPLM_A and OPLM_B are the item parameters used in PMS. BILOG_A and BILOG_B denote the parameters obtained from multiple choice item data. These calibrations are used in live CAT administration phase. Finally, MIRT_A and MIRT_B give the item parameters which are calibrated by MIRT and used in simulation phases of the current study. Parameter names endingsss up with _A stands for item discrimination parameter and _B stands for item discrimination parameter.

Table A.1 Item Parameters Calibrated by BILOG-MG, OPLM and MIRT

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 1	0,623	-1,218	2	-0,332	0,457	-2,518
Item 2	0,563	0,768	3	0,667	0,459	0,885
Item 3	0,853	-2,151	4	-0,379	0,614	-2,831
Item 4	0,834	-0,723				
Item 5	0,374	0,888	1	1,231	0,377	1,412
Item 6	0,783	1,370	4	0,737	0,643	1,443
Item 7	0,882	-1,407	4	-0,082	0,787	-1,420
Item 8	1,056	-2,302	5	-0,286	1,054	-1,911
Item 9	0,384	-0,328	2	0,148	0,293	-0,853
Item 10	1,011	-1,792	4	-0,172	0,752	-2,091
Item 11	0,758	-0,546	3	0,167	0,678	-0,693
Item 12	0,837	-1,549	4	-0,112	0,657	-1,602
Item 13	0,937	-1,448	4	-0,094	0,832	-1,630
Item 14	0,861	-1,533	4	-0,160	0,962	-1,362
Item 15	0,870	-1,139	4	-0,009	0,736	-1,063
Item 16	1,089	-1,712	5	-0,094	1,069	-1,014
Item 17	0,384	-0,425	2	0,020	0,364	-1,321
Item 18	0,824	-1,541	4	-0,128	0,804	-1,684
Item 19	0,746	-0,893	3	-0,029	0,521	-1,471
Item 20	0,982	-1,634	5	-0,081	0,913	-1,676
Item 21	0,712	-1,855	3	-0,386	0,528	-2,280
Item 22	1,088	-1,930	4	-0,204	0,826	-1,569
Item 23	0,807	-1,790	3	-0,341	0,762	-1,904
Item 24	0,972	-0,880	5	0,105	0,930	-0,826
Item 25	1,079	-2,705	6	-0,346	0,694	-2,821
Item 26	0,678	0,273	3	0,408	0,532	-0,298
Item 27	0,708	0,528	3	0,493	0,591	0,535
Item 28	0,665	0,910				
Item 29	0,671	-0,402	3	0,152	0,610	-0,813
Item 30	0,983	0,205				
Item 31	0,270	1,234	1	1,486	0,241	2,797
Item 32	0,573	-0,596	2	0,013	0,399	-0,894
Item 33	0,560	-0,435	3	0,140	0,417	-0,780
Item 34	0,490	-0,413	2	0,138	0,303	-0,780
Item 35	0,508	0,847	2	0,731	0,432	0,700
Item 36	0,508	-0,839	2	-0,410	0,357	-1,702

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 37	0,628	-0,814				
Item 38	0,857	0,539				
Item 39	1,025	1,274				
Item 40	0,974	2,961				
Item 41	0,839	0,477				
Item 42	0,974	2,340				
Item 43	1,062	3,872				
Item 44	0,877	-0,803	4	-0,159	0,657	-1,271
Item 45	0,812	-0,820				
Item 46	0,841	-1,258				
Item 47	0,577	1,407	3	0,483	0,513	1,532
Item 48	0,864	-0,455				
Item 49	0,718	-0,428	3	-0,149	0,560	-0,507
Item 50	0,844	2,017				
Item 51	1,013	0,883				
Item 52	0,921	0,937				
Item 53	0,436	0,572	2	0,316	0,402	0,151
Item 54	0,858	-0,982	5	-0,215	0,691	-1,119
Item 55	0,635	-1,030	3	-0,353	0,565	-1,237
Item 56	0,604	-1,173	3	-0,447	0,514	-1,700
Item 57	0,612	-1,294	4	-0,375	0,688	-1,476
Item 58	0,587	-1,351	2	-0,741	0,486	-2,101
Item 59	0,741	-0,287	3	-0,056	0,520	-0,349
Item 60	0,736	-0,436				
Item 61	0,745	-0,877				
Item 62	1,097	0,441				
Item 63	0,709	0,083				
Item 64	0,910	0,302				
Item 65	0,557	-0,679	2	-0,309	0,441	-0,988
Item 66	0,755	0,056				
Item 67	0,819	0,921				
Item 68	0,764	-1,193				
Item 69	0,726	-0,740	2	-0,377	0,810	-1,172
Item 70	0,537	0,260	2	0,111	0,442	-0,239
Item 71	0,440	-0,532	1	-0,614	0,395	-1,342
Item 72	0,819	-0,005	4	0,044	0,840	-0,144
Item 73	0,716	-0,840	3	-0,260	0,603	-1,469
Item 74	1,057	-1,496				

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 75	0,974	2,174				
Item 76	0,625	0,034	3	0,019	0,462	-0,409
Item 77	1,171	-2,458				
Item 78	1,066	-0,602				
Item 79	0,661	-0,135	3	0,039	0,558	-0,632
Item 80	1,047	0,312				
Item 81	0,928	0,956				
Item 82	0,777	0,997				
Item 83	0,407	-0,518	1	-0,423	0,338	-1,081
Item 84	0,445	1,080				
Item 85	0,755	-1,361	2	-0,552	0,481	-1,585
Item 86	0,709	-0,684	3	-0,117	0,653	-1,245
Item 87	0,717	0,135	3	0,208	0,602	-0,237
Item 88	0,854	0,793	4	0,369	0,985	-0,196
Item 89	0,753	-1,049	3	-0,272	0,629	-1,805
Item 90	0,638	-1,313	2	-0,551	0,561	-1,734
Item 91	0,946	0,497				
Item 92	0,717	1,111				
Item 93	1,052	-1,556	4	-0,309	1,253	-1,460
Item 94	1,017	0,908				
Item 95	0,709	-0,635	3	-0,105	0,580	-0,629
Item 96	1,006	-0,542	3	-0,024	0,823	-0,716
Item 97	1,155	0,930				
Item 98	0,764	0,422	3	0,268	0,589	0,140
Item 99	0,643	-0,340	2	-0,061	0,495	-0,605
Item 100	0,760	-1,097	3	-0,267	0,566	-1,434
Item 101	1,111	-2,545	4	-0,638	0,965	-2,124
Item 102	0,821	-1,758	2	-0,793	0,575	-2,181
Item 103	0,802	-1,646	3	-0,518	0,732	-1,698
Item 104	1,135	-1,991	4	-0,430	0,838	-1,654
Item 105	1,074	-2,312	3	-0,714	0,705	-2,473
Item 106	1,023	-2,009	4	-0,454	0,799	-1,780
Item 107	0,983	-1,544	3	-0,395	0,778	-1,809
Item 108	1,134	-0,331				
Item 109	0,959	-1,347	4	-0,241	0,719	-1,398
Item 110	0,663	-0,601	2	-0,235	0,688	-1,382
Item 111	0,996	-0,609	4	0,002	0,766	-0,780
Item 112	0,463	-0,749	2	-0,276	0,393	-1,475

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 113	0,782	-0,223	3	0,075	0,748	-0,473
Item 114	0,405	-0,970	1	-0,880	0,265	-2,344
Item 115	0,746	1,184	3	0,532	0,795	0,021
Item 116	0,509	-0,239	2	-0,005	0,435	-0,565
Item 117	0,629	-0,101				
Item 118	0,578	-0,238	2	0,005	0,606	-0,493
Item 119	1,055	-1,758	4	-0,359	1,083	-1,517
Item 120	0,906	-1,823	4	-0,414	0,841	-1,665
Item 121	0,910	-0,570	4	-0,038	0,905	-0,652
Item 122	0,887	-0,505	4	-0,004	0,841	-0,702
Item 123	0,793	-1,624	3	-0,506	0,626	-1,923
Item 124	0,729	-2,172	3	-0,814	0,647	-2,362
Item 125	0,733	-0,331	3	0,013	0,557	-0,439
Item 126	0,847	-1,779				
Item 127	0,910	-1,313	4	-0,230	0,756	-1,305
Item 128	0,953	-1,162	4	-0,184	0,756	-1,091
Item 129	0,640	-0,221	2	0,009	0,448	-0,455
Item 130	0,834	-1,299	2	-0,573	0,598	-1,623
Item 131	0,653	0,823	3	0,425	0,608	0,795
Item 132	0,957	-1,496	4	-0,301	1,003	-1,649
Item 133	0,942	-1,326	3	-0,313	0,571	-1,345
Item 134	0,848	-2,063	3	-0,591	0,549	-1,919
Item 135	0,516	-1,184	3	-0,274	0,527	-1,860
Item 136	0,470	-1,065	2	-0,363	0,287	-2,863
Item 137	0,763	-1,490	2	-0,473	0,442	-2,758
Item 138	0,695	-0,730	3	-0,005	0,690	-0,806
Item 139	0,627	-1,085	2	-0,408	0,323	-2,446
Item 140	0,563	-0,423	2	0,005	0,477	-0,825
Item 141	0,294	0,361				
Item 142	0,425	-2,115	4	-0,332		
Item 143	0,809	-2,050	6	-0,233	0,859	-2,646
Item 144	0,355	-0,174	2	0,171	0,281	-0,229
Item 145	0,343	-1,104	2	-0,311		
Item 146	0,543	-1,007	3	-0,153	0,554	-1,405
Item 147	0,836	-1,982	5	-0,262	0,709	-2,263
Item 148	0,468	-0,166	3	0,096	0,501	-0,509
Item 149	0,932	-1,606	5	-0,097	0,592	-2,011
Item 150	0,309	-0,066	2	0,144	0,341	-0,203

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 151	0,632	-1,976	5	-0,262	0,501	-3,077
Item 152	0,426	-0,517	2	-0,093	0,532	-0,745
Item 153	0,440	0,031	3	0,238		
Item 154	0,260	-1,677	3	-0,374		
Item 155	0,384	0,029	2	0,211	0,431	-0,301
Item 156	0,521	-0,742	4	-0,005	0,568	-1,198
Item 157	0,270	-2,440	1	-2,137		
Item 158	0,811	-2,105	5	-0,393	0,463	-3,127
Item 159	0,940	-2,579	4	-0,634	0,825	-2,917
Item 160	0,536	0,001	3	0,247	0,545	0,010
Item 161	0,827	-1,180	5	-0,026	0,915	-1,169
Item 162	0,964	-1,218				
Item 163	0,268	0,578	1	0,427	0,285	1,022
Item 164	0,414	-1,407				
Item 165	0,731	-0,087				
Item 166	0,679	-0,216				
Item 167	0,768	-1,370				
Item 168	0,809	-1,278				
Item 169	0,724	-1,188	4	-0,435		
Item 170	0,503	0,174				
Item 171	0,788	0,637				
Item 172	0,347	-1,731	2	-0,964		
Item 173	0,548	-0,542	4	-0,246		
Item 174	0,561	-0,914	4	-0,337		
Item 175	0,634	-0,389	4	-0,174	0,795	-0,718
Item 176	0,767	0,062	4	0,069	0,566	0,205
Item 177	0,762	0,566				
Item 178	0,730	0,563				
Item 179	0,297	-0,018	1	-0,234	0,202	-0,676
Item 180	0,611	0,737	3	0,273	0,574	0,679
Item 181	0,471	1,380	2	0,710	0,340	2,367
Item 182	0,801	-0,213				
Item 183	0,549	-0,094	2	0,012	0,527	0,011
Item 184	0,642	-0,748	3	-0,184	0,778	-0,839
Item 185	0,273	0,844	1	0,871	0,281	1,776
Item 186	0,581	-0,143	3	0,056	0,516	-0,392
Item 187	0,425	-0,145	2	-0,020	0,324	-0,737
Item 188	0,541	-0,373	3	-0,042	0,635	-0,572

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 189	0,626	-0,012	3	0,117	0,666	-0,128
Item 190	0,533	0,179	2	0,191	0,470	-0,126
Item 191	0,712	-0,990				
Item 192	0,644	-0,754	3	-0,200	0,599	-1,210
Item 193	0,934	-1,480	4	-0,314	1,157	-1,091
Item 194	0,761	-1,194	3	-0,361	0,624	-1,559
Item 195	0,737	-0,496				
Item 196	0,637	-0,363	3	0,017	0,724	-0,453
Item 197	0,697	-0,312				
Item 198	0,498	0,003				
Item 199	0,713	-1,001				
Item 200	0,544	-1,560	2	-0,910	0,836	-2,305
Item 201	0,682	-2,605	2	-1,613	0,535	-2,902
Item 202	0,743	-0,461				
Item 203	0,436	-0,875	2	-0,530	0,507	-1,694
Item 204	0,962	0,404				
Item 205	1,183	-1,848	4	-0,378	1,297	-1,159
Item 206	0,775	0,023				
Item 207	0,820	-0,064	4	0,122	0,731	0,039
Item 208	0,976	-1,802	3	-0,597	0,492	-2,176
Item 209	1,287	-3,015	3	-1,189	0,919	-2,970
Item 210	0,656	0,838				
Item 211	0,769	-0,997				
Item 212	0,768	-1,122				
Item 213	0,706	-1,623	3	-0,663	0,956	-1,901
Item 214	0,643	-0,342	4	-0,114	0,919	-0,809
Item 215	0,619	-0,269				
Item 216	0,905	-3,179	4	-1,178	0,729	-3,025
Item 217	0,791	-0,042	3	0,101	0,598	-0,099
Item 218	0,748	-0,843				
Item 219	0,559	-0,482	2	-0,180	0,434	-1,027
Item 220	0,748	-0,186	3	0,047	0,637	-0,231
Item 221	0,710	-0,650				
Item 222	0,857	-1,193	4	-0,253	0,777	-1,216
Item 223	0,559	0,196	2	0,181	0,591	-0,074
Item 224	0,827	-1,796	3	-0,682	0,718	-2,137
Item 225	0,420	0,716	2	0,719	0,303	0,980
Item 226	0,962	0,894				

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 227	0,574	0,773	2	0,711	0,463	0,743
Item 228	0,449	0,710	1	0,938	0,307	0,802
Item 229	0,884	-0,773	4	0,130	0,639	-1,009
Item 230	0,947	0,921	3	0,645	0,882	0,456
Item 231	0,727	-0,292	3	0,237	0,583	-0,971
Item 232	0,557	-0,754				
Item 233	0,637	-0,706	3	0,067	0,493	-1,230
Item 234	0,687	-0,685	3	0,046	0,444	-1,456
Item 235	0,424	-0,227	1	-0,124	0,306	-0,829
Item 236	0,677	-0,479	2	0,042	0,444	-1,138
Item 237	0,862	-0,789	4	0,107	0,654	-1,118
Item 238	0,398	1,997	1	2,226	0,430	2,543
Item 239	0,503	0,866	2	0,747	0,350	1,246
Item 240	1,133	-2,039	6	-0,137	1,057	-1,988
Item 241	0,975	-0,418				
Item 242	0,444	0,897	2	0,780	0,282	1,761
Item 243	0,655	-1,212	2	-0,544	0,269	-2,814
Item 244	0,520	0,684	2	0,667	0,367	0,258
Item 245	0,471	0,538	2	0,608	0,356	0,671
Item 246	0,867	-0,572	3	0,129	0,556	-1,263
Item 247	0,952	-1,152				
Item 248	0,797	1,130	3	0,528	0,618	0,886
Item 249	0,460	-0,029	2	0,096	0,479	-0,634
Item 250	0,771	0,337	3	0,288	0,784	-0,105
Item 251	0,945	0,388				
Item 252	0,795	1,742				
Item 253	0,281	0,717				
Item 254	0,253	1,309			0,153	3,387
Item 255	0,696	0,643	3	0,369	0,507	0,252
Item 256	0,506	0,791	2	0,541	0,389	0,841
Item 257	0,921	0,597	3	0,321	0,609	0,345
Item 258	0,769	-0,172	3	0,069	0,585	-0,797
Item 259	0,888	-0,898	3	-0,204	0,657	-0,944
Item 260	0,977	0,090				
Item 261	0,703	0,759	3	0,403	0,527	0,348
Item 262	1,007	-0,769	4	-0,091	0,807	-0,893
Item 263	0,968	0,169	3	0,137	0,581	-0,020
Item 264	0,772	-0,417	3	-0,039	0,500	-1,057

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 265	0,853	2,174				
Item 266	0,354	0,749	1	0,847	0,355	0,877
Item 267	1,011	-0,426				
Item 268	1,214	-0,418	5	0,122	1,164	-0,596
Item 269	0,698	-0,099	2	0,101	0,519	-0,873
Item 270	1,099	-1,329	5	-0,188	1,080	-1,536
Item 271	0,640	0,203	2	0,301	0,522	-0,106
Item 272	0,891	-1,050	4	-0,172	0,698	-1,422
Item 273	0,732	-0,242	2	0,022	0,647	-0,590
Item 274	0,731	1,890				
Item 275	0,985	-1,280	4	-0,247	0,756	-1,796
Item 276	0,842	0,127	3	0,250	0,657	-0,122
Item 277	0,468	1,107	1	1,206	0,487	1,075
Item 278	0,959	-0,643	4	0,000	0,930	-0,829
Item 279	0,488	1,137	2	0,806	0,435	1,018
Item 280	0,725	0,448	3	0,381	0,563	-0,150
Item 281	0,708	0,328	2	0,361	0,588	0,242
Item 282	1,063	0,801				
Item 283	0,785	0,085	3	0,246	0,583	-0,278
Item 284	0,798	-1,375	3	-0,451	0,636	-2,168
Item 285	1,008	-0,689	4	0,012	0,869	-0,779
Item 286	0,472	0,144				
Item 287	1,189	-0,210				
Item 288	0,812	-0,493	2	-0,086	0,594	-1,015
Item 289	0,833	-0,164	3	0,157	0,619	-0,349
Item 290	1,273	-0,868	5	0,001	0,846	-1,011
Item 291	0,838	0,459	2	0,417	0,534	0,136
Item 292	1,022	-0,746	4	-0,021	0,733	-0,957
Item 293	0,569	0,127	2	0,233	0,381	-0,295
Item 294	0,440	0,737				
Item 295	0,936	0,440	4	0,366	0,666	0,289
Item 296	0,888	-0,709	3	-0,087	0,566	-1,447
Item 297	0,911	-1,069				
Item 298	0,482	0,482	2	0,441	0,373	0,493
Item 299	0,776	-0,152	3	0,139	0,551	-0,589
Item 300	1,000	-0,173	3	0,138	0,643	-0,377
Item 301	0,863	-0,580	3	-0,041	0,480	-0,633
Item 302	0,848	-0,584	3	-0,011	0,663	-1,052

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 303	0,402	1,741	1	1,871	0,360	2,390
Item 304	0,885	0,401	3	0,362	0,725	0,313
Item 305	0,634	0,715				
Item 306	1,218	-0,430	5	0,127	1,249	-0,810
Item 307	0,504	0,172	1	0,227	0,358	-0,625
Item 308	0,924	0,227				
Item 309	1,089	-1,092	3	-0,039	0,815	-1,448
Item 310	1,184	-2,447	4	-0,310	0,897	-2,026
Item 311	0,898	0,453	3	0,488	0,619	0,429
Item 312	0,737	-1,974	2	-0,746	0,465	-2,907
Item 313	1,146	-1,265	3	-0,105	0,742	-1,399
Item 314	0,754	0,844	3	0,656	0,494	0,923
Item 315	0,362	2,088	1	2,356	0,350	2,502
Item 316	0,951	1,472				
Item 317	0,347	0,663	1	1,013	0,342	1,150
Item 318	0,667	-0,331	2	0,164	0,523	-0,689
Item 319	0,587	0,307				
Item 320	0,671	0,558	2	0,652	0,486	0,427
Item 321	0,488	0,893	2	0,816	0,440	0,824
Item 322	0,593	-0,433	3	0,237	0,439	-0,891
Item 323	0,833	-1,272	2	-0,383	0,568	-2,063
Item 324	1,155	-1,983	5	-0,099	0,818	-1,916
Item 325	1,225	0,151				
Item 326	0,519	0,894	2	0,834	0,368	0,845
Item 327	0,625	-0,349	2	0,142	0,586	-0,648
Item 328	0,668	-0,523	3	0,179	0,461	-0,800
Item 329	0,431	-0,581	1	-0,398	0,265	-1,424
Item 330	0,837	0,236				
Item 331	0,689	3,060				
Item 332	1,235	1,226				
Item 333	0,348	-0,257				
Item 334	0,500	0,169	1	0,150	0,389	-0,730
Item 335	1,119	0,357	4	0,227	0,923	-0,586
Item 336	0,644	2,924				
Item 337	0,816	-0,566				
Item 338	0,887	-0,368	3	0,028	0,607	-0,796
Item 339	1,103	2,269				
Item 340	0,539	0,430	3	0,344	0,501	0,270

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 341	0,717	1,348				
Item 342	1,046	1,257	4	0,475	0,754	0,805
Item 343	0,567	-0,562				
Item 344	0,905	0,549				
Item 345	1,432	2,821				
Item 346	0,969	1,698	3	0,674	0,684	1,056
Item 347	1,355	3,520				
Item 348	0,924	-0,057	3	0,141	0,665	-0,198
Item 349	0,864	0,532	3	0,324	0,647	-0,194
Item 350	0,876	1,450				
Item 351	0,751	0,224	3	0,308	0,540	-0,011
Item 352	0,796	0,580	2	0,504	0,556	0,204
Item 353	0,553	-0,028	2	0,206	0,373	-0,288
Item 354	1,074	-0,383	4	0,115	0,797	-0,610
Item 355	0,297	1,134	1	1,359	0,234	2,703
Item 356	0,888	0,899	3	0,559	0,763	-0,088
Item 357	1,148	0,030	4	0,265	0,901	-0,622
Item 358	1,140	0,985	5	0,461	0,842	0,537
Item 359	0,842	-0,129	2	0,152	0,590	-0,769
Item 360	0,740	0,798	3	0,521	0,595	0,531
Item 361	1,039	0,823	5	0,429	0,904	0,300
Item 362	1,119	0,294	4	0,329	0,837	-0,804
Item 363	1,189	0,263				
Item 364	0,713	-1,126	2	-0,370	0,478	-1,667
Item 365	1,178	0,344	4	0,346	0,977	-0,395
Item 366	0,664	1,461	2	0,953	0,534	1,434
Item 367	0,591	-0,873	2	-0,242	0,432	-1,445
Item 368	0,745	-0,334	2	0,071	0,630	-0,976
Item 369	0,274	0,858	1	1,090	0,209	2,185
Item 370	0,645	1,157				
Item 371	0,906	0,063	4	0,249	0,761	-0,193
Item 372	1,256	-0,035	4	0,242	0,860	-0,164
Item 373	0,700	-0,184	2	0,118	0,547	-0,503
Item 374	1,272	-0,342	4	0,159	1,070	-0,620
Item 375	1,151	0,158	4	0,298	0,878	-0,477
Item 376	0,846	0,662	3	0,460	0,730	0,496
Item 377	1,375	1,293				
Item 378	0,762	-0,579	3	0,017	0,666	-0,750

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 379	1,259	-0,972	5	0,004	1,023	-0,733
Item 380	1,385	-1,000	6	0,024	1,272	-0,821
Item 381	0,372	0,617	1	0,810	0,244	1,436
Item 382	1,007	0,068	4	0,247	0,772	-0,330
Item 383	1,232	-2,126	4	-0,335	0,981	-1,821
Item 384	1,159	0,119	4	0,289	0,780	-0,176
Item 385	0,251	1,854	1	2,120	0,275	3,112
Item 386	0,472	1,427	2	1,002	0,393	1,947
Item 387	0,617	0,803	2	0,672	0,505	0,805
Item 388	1,632	-1,753	7	-0,073	1,314	-1,542
Item 389	1,392	-0,388	4	0,210	0,797	-0,766
Item 390	1,345	-0,177	4	0,279	0,694	-0,568
Item 391	1,185	-0,845	4	0,062	0,772	-1,257
Item 392	0,893	1,746	3	0,947	0,566	1,300
Item 393	1,401	-0,906	5	0,091	0,952	-1,146
Item 394	0,844	1,450	3	0,842	0,557	1,226
Item 395	0,646	0,578	1	0,660	0,394	0,242
Item 396	1,101	-0,171	3	0,249	0,796	-0,385
Item 397	0,882	-0,227	2	0,115	0,521	-0,678
Item 398	0,746	0,048	2	0,310	0,329	-0,440
Item 399	1,267	0,129				
Item 400	1,191	-0,064	4	0,332	0,883	-0,600
Item 401	0,903	0,145	2	0,329	0,502	-0,290
Item 402	0,686	0,403	2	0,496	0,319	0,187
Item 403	1,267	0,376	4	0,465	0,849	0,026
Item 404	1,128	0,538	5	0,484	0,879	-0,020
Item 405	1,148	-0,165	3	0,249	0,602	-0,778
Item 406	0,550	1,934	1	2,073	0,303	3,072
Item 407	1,526	0,511	8	0,473	0,989	-0,131
Item 408	1,141	-0,186	5	0,291	0,785	-0,630
Item 409	1,736	-1,172				
Item 410	0,791	0,921	3	0,654	0,531	0,511
Item 411	1,292	-0,460	4	0,190	0,754	-0,917
Item 412	1,296	-0,466	6	0,223	1,499	-0,864
Item 413	1,497	-0,667	7	0,205	1,729	-0,906
Item 414	0,601	0,396	1	0,437	0,339	-0,136
Item 415	0,723	0,273	2	0,379	0,359	-0,132
Item 416	0,795	1,355	3	0,789	0,484	0,775

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 417	0,939	1,714	3	0,908	0,685	1,060
Item 418	0,907	0,308	3	0,389	0,538	-0,387
Item 419	1,507	-0,038	7	0,338	1,088	-0,384
Item 420	0,448	1,182	1	1,107	0,284	1,964
Item 421	1,083	-0,552	3	-0,183	0,540	-1,669
Item 422	1,713	3,602				
Item 423	1,501	-1,054				
Item 424	0,767	0,872				
Item 425	0,841	0,181				
Item 426	0,381	0,970			0,122	3,201
Item 427	0,896	1,597	3	0,613	0,519	1,511
Item 428	0,746	1,679	3	0,674	0,440	1,723
Item 429	0,520	0,962	1	0,826	0,303	1,245
Item 430	0,595	1,004				
Item 431	1,760	-0,912				
Item 432	0,844	-0,317	2	-0,222	0,333	-1,386
Item 433	0,593	0,804	2	0,445	0,345	0,827
Item 434	0,730	2,388	2	1,204	0,428	3,035
Item 435	1,346	1,108				
Item 436	1,269	3,124				
Item 437	0,863	0,476				
Item 438	0,539	1,923	2	1,033	0,384	2,711
Item 439	0,897	0,853				
Item 440	0,944	0,607				
Item 441	0,330	1,771	1	1,715	0,202	3,121
Item 442	1,192	0,563				
Item 443	0,279	1,162			0,167	3,314
Item 444	1,421	-0,943	5	-0,190	0,831	-1,590
Item 445	0,992	0,643				
Item 446	1,192	0,314	3	0,157	0,625	-0,099
Item 447	1,463	-0,396				
Item 448	1,382	0,962	5	0,312	0,824	0,380
Item 449	0,618	0,321	1	0,127	0,272	-0,457
Item 450	0,911	1,101	3	0,455	0,591	0,653
Item 451	1,592	-0,867				
Item 452	1,133	0,882				
Item 453	1,129	0,793				
Item 454	1,796	3,327				

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 455	0,678	2,221	2	1,165	0,427	2,790
Item 456	0,479	2,678	1	2,578	0,315	3,175
Item 457	1,211	1,230				
Item 458	1,123	-0,592				
Item 459	0,782	0,172	2	0,092	0,498	-0,373
Item 460	1,735	0,129				
Item 461	1,372	0,279				
Item 462	1,277	-0,029	4	0,279	0,708	-0,633
Item 463	0,970	0,321				
Item 464	1,103	-0,551	3	0,038	0,502	-1,605
Item 465	1,156	-0,438				
Item 466	0,678	0,829	2	0,693	0,358	0,745
Item 467	0,858	0,509	2	0,489	0,534	-0,182
Item 468	0,606	-0,413	1	-0,485	0,238	-2,687
Item 469	0,850	1,875	3	0,994	0,617	1,789
Item 470	1,100	-0,630	2	-0,161	0,596	-1,382
Item 471	1,258	0,379				
Item 472	1,477	-0,247				
Item 473	0,984	0,781				
Item 474	0,876	0,229				
Item 475	1,380	-0,568	4	0,114	0,682	-1,040
Item 476	0,723	0,097	1	0,129	0,421	-0,615
Item 477	1,298	-0,603	3	0,031	0,646	-1,301
Item 478	0,821	-0,336	2	0,010	0,387	-1,388
Item 479	1,102	0,610	3	0,491	0,685	-0,175
Item 480	0,827	-0,551				
Item 481	1,216	1,227				
Item 482	1,147	3,041				
Item 483	1,244	3,575				
Item 484	0,685	1,004	2	0,781	0,417	0,955
Item 485	1,001	-0,804	2	-0,297	0,569	-1,850
Item 486	0,986	1,094	3	0,681	0,510	0,595
Item 487	1,474	-1,411	3	-0,346	0,582	-2,387
Item 488	0,846	0,551	2	0,511	0,422	0,026
Item 489	1,403	-1,099	4	-0,078	0,678	-1,634
Item 490	0,972	0,397	3	0,423	0,499	-0,090
Item 491	1,454	3,181				
Item 492	1,029	0,986	3	0,637	0,556	0,565

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 493	0,929	2,033	3	1,029	0,582	1,804
Item 494	1,350	-1,073	3	-0,180	0,681	-1,873
Item 495	0,474	1,278				
Item 496	1,078	-0,049	4	-0,256	0,785	-0,403
Item 497	1,055	0,180	4	-0,179	0,747	-0,195
Item 498	0,704	0,364	2	-0,112	0,331	-0,393
Item 499	1,479	-1,124	6	-0,510	0,976	-1,433
Item 500	1,009	2,990				
Item 501	1,089	0,259	3	-0,167	0,740	-0,204
Item 502	0,707	0,276	2	-0,131	0,462	0,011
Item 503	0,942	1,028				
Item 504	1,482	-0,795				
Item 505	1,375	-1,170	4	-0,617	0,805	-1,599
Item 506	0,592	0,517	1	0,056	0,328	0,533
Item 507	0,880	0,464	2	-0,067	0,520	-0,095
Item 508	0,927	0,613	3	-0,016	0,766	0,042
Item 509	0,670	0,994				
Item 510	0,679	1,643				
Item 511	0,814	0,385	2	-0,110	0,459	-0,180
Item 512	1,289	0,073				
Item 513	1,049	0,377	3	-0,133	0,646	-0,343
Item 514	0,799	0,513	3	-0,070	0,674	-0,091
Item 515	0,635	0,064	2	-0,174	0,542	-0,292
Item 516	0,571	0,058	2	-0,199	0,443	-0,293
Item 517	1,138	-2,037	4	-0,798	0,905	-1,951
Item 518	1,088	-1,637	4	-0,683	0,831	-1,601
Item 519	0,734	-0,692	3	-0,440	0,585	-1,078
Item 520	0,826	-0,105	2	-0,394	0,385	-0,729
Item 521	0,968	-0,372	3	-0,428	0,617	-0,959
Item 522	1,080	-0,065	3	-0,288	0,556	-0,689
Item 523	1,408	-0,494	5	-0,343	0,764	-0,872
Item 524	1,001	0,698				
Item 525	0,643	1,702	2	0,624	0,424	2,081
Item 526	0,633	0,451	2	-0,029	0,330	0,136
Item 527	1,058	2,345				
Item 528	1,479	-0,533				
Item 529	0,991	-0,265	3	-0,367	0,995	-0,646
Item 530	0,684	-0,330	2	-0,384	0,568	-0,889

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 531	0,613	-0,275	3	-0,392	0,656	-1,178
Item 532	0,719	-0,687	3	-0,513	0,589	-1,286
Item 533	0,991	-1,613	3	-0,796	0,719	-1,881
Item 534	0,722	-0,533	3	-0,371	0,626	-0,757
Item 535	1,203	-0,436	4	-0,402	0,719	-1,163
Item 536	1,102	0,444	5	-0,112	0,934	-0,271
Item 537	0,800	0,232	2	-0,215	0,388	-0,408
Item 538	0,812	1,898				
Item 539	0,865	-0,306				
Item 540	0,954	1,203	3	0,164	0,626	0,587
Item 541	0,838	0,511				
Item 542	1,120	0,521				
Item 543	1,398	-0,433	6	0,291	1,012	-1,042
Item 544	1,063	-0,525	4	0,231	0,815	-1,149
Item 545	0,747	1,712	3	1,052	0,444	1,635
Item 546	1,404	-0,591	5	0,226	0,961	-1,225
Item 547	1,138	-0,792	3	0,131	0,663	-1,418
Item 548	0,988	1,000	4	0,722	0,672	0,201
Item 549	0,922	-0,337	3	0,304	0,608	-1,112
Item 550	1,051	0,676	4	0,626	0,745	0,126
Item 551	1,046	0,415	4	0,556	0,676	-0,286
Item 552	0,985	1,312	3	0,864	0,621	0,520
Item 553	0,983	0,444	3	0,572	0,635	-0,038
Item 554	1,666	0,523	6	0,537	0,981	-0,366
Item 555	0,732	-0,114	2	0,261	0,554	-0,826
Item 556	1,389	-1,185	4	0,112	0,874	-1,518
Item 557	1,117	-0,280	4	0,324	0,792	-0,992
Item 558	0,607	-0,284	2	0,204	0,456	-1,080
Item 559	1,309	-0,097	5	0,364	0,977	-0,666
Item 560	1,480	-1,369	5	0,050	1,159	-1,514
Item 561	1,353	-0,264	5	0,292	1,070	-0,917
Item 562	1,702	0,123	7	0,403	1,233	-0,760
Item 563	1,240	-0,003	4	0,424	0,809	-0,624
Item 564	1,254	0,147	4	0,463	0,878	-0,494
Item 565	0,725	1,680	2	1,242	0,452	1,365
Item 566	1,212	0,587	4	0,584	0,898	-0,170
Item 567	0,695	-0,124	2	0,340	0,454	-0,785
Item 568	0,727	0,291	2	0,545	0,439	-0,402

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 569	1,012	0,596	3	0,620	0,786	-0,073
Item 570	1,638	-0,903	5	0,224	1,049	-0,933
Item 571	0,265	1,749	1	2,186	0,269	3,401
Item 572	0,830	0,757	3	0,694	0,589	0,117
Item 573	1,397	-0,679	7	0,190	1,495	-1,180
Item 574	1,183	0,103	4	0,344	0,792	-0,813
Item 575	0,892	0,988	2	0,700	0,621	0,075
Item 576	1,153	-0,375	4	0,166	0,777	-1,204
Item 577	0,255	0,298	1	0,705	0,212	0,420
Item 578	1,604	0,247	7	0,487	1,124	-0,439
Item 579	1,657	-1,025				
Item 580	1,149	1,478				
Item 581	1,828	3,291				
Item 582	1,801	3,668				
Item 583	1,120	0,594				
Item 584	1,642	1,235				
Item 585	0,345	-1,063	1	-1,019	0,251	-3,139
Item 586	1,323	-0,042	4	0,129	0,868	-0,543
Item 587	1,336	-0,124				
Item 588	1,424	2,146				
Item 589	0,271	1,468			0,156	3,350
Item 590	1,155	0,391	4	0,258	0,788	-0,407
Item 591	1,032	1,778				
Item 592	1,978	3,708				
Item 593	1,345	-0,982	4	-0,115	0,757	-1,517
Item 594	0,477	1,424				
Item 595	1,211	-0,211				
Item 596	1,642	-0,679				
Item 597	0,795	1,531	3	0,702	0,568	0,939
Item 598	1,500	3,952				
Item 599	0,957	1,037				
Item 600	0,888	2,770				
Item 601	1,803	2,961				
Item 602	0,664	1,587	2	0,931	0,419	1,883
Item 603	0,928	2,296				
Item 604	0,870	1,045				
Item 605	1,304	3,133				
Item 606	1,462	-0,510	5	0,050	0,932	-1,028

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 607	0,400	0,426			0,170	0,317
Item 608	1,043	0,940				
Item 609	1,630	1,797				
Item 610	1,136	0,483				
Item 611	1,003	-0,687	3	-0,119	0,638	-1,509
Item 612	0,899	0,523	2	0,477	0,594	-0,014
Item 613	1,593	0,168	4	0,292	1,181	-0,475
Item 614	1,261	0,605	3	0,438	1,024	-0,158
Item 615	1,207	0,660	4	0,458	0,912	-0,074
Item 616	0,691	3,556				
Item 617	1,638	0,188	5	0,306	1,066	-0,545
Item 618	0,685	0,578	2	0,524	0,449	0,205
Item 619	1,271	1,012	4	0,546	0,719	0,196
Item 620	1,109	-0,233	3	0,173	0,728	-0,821
Item 621	0,921	0,411	3	0,422	0,528	-0,065
Item 622	0,674	1,263				
Item 623	1,352	-0,501				
Item 624	1,701	-0,857				
Item 625	1,540	0,328				
Item 626	1,455	-1,458	4	-0,125	0,817	-1,782
Item 627	1,398	2,502				
Item 628	1,496	2,191				
Item 629	1,419	0,152	4	0,314	1,029	-0,499
Item 630	1,711	1,008	5	0,496	1,083	-0,018
Item 631	1,506	2,537				
Item 632	0,820	1,135				
Item 633	1,612	1,378				
Item 634	1,759	2,563				
Item 635	1,097	-0,710	3	0,013	0,630	-1,582
Item 636	2,116	-0,926	7	0,114	1,290	-0,984
Item 637	1,508	-0,938				
Item 638	0,899	0,080	2	0,276	0,616	-0,586
Item 639	1,261	3,321				
Item 640	0,878	-0,187	2	0,147	0,534	-0,760
Item 641	0,741	0,954	2	0,753	0,460	0,752
Item 642	1,227	-0,100	3	0,227	0,679	-0,612
Item 643	1,059	0,865	3	0,578	0,644	0,427
Item 644	0,528	1,697	2	1,215	0,364	2,347

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 645	1,114	-0,284	3	0,172	0,717	-0,725
Item 646	1,052	0,394	3	0,416	0,648	-0,253
Item 647	1,658	-0,354	5	0,201	0,904	-0,804
Item 648	1,198	-0,224	3	0,187	0,671	-0,721
Item 649	1,447	-1,235				
Item 650	1,433	3,851				
Item 651	0,341	0,056	1	0,304	0,247	-0,346
Item 652	1,009	0,135	3	0,328	0,688	-0,363
Item 653	0,514	1,489	2	1,108	0,349	2,017
Item 654	1,653	0,195	5	0,320	1,083	-0,485
Item 655	1,458	1,196	4	0,581	0,829	0,334
Item 656	1,318	-0,611	4	0,117	0,809	-0,904
Item 657	1,670	1,479				
Item 658	0,809	0,384	2	0,445	0,473	-0,032
Item 659	0,393	1,389	1	1,540	0,299	2,463
Item 660	1,927	0,226	6	0,319	1,337	-0,516
Item 661	1,802	-0,880	6	0,116	1,127	-1,023
Item 662	0,983	0,629	3	0,506	0,631	-0,042
Item 663	1,900	-1,806				
Item 664	1,083	-0,074	3	0,246	0,693	-0,590
Item 665	1,839	-0,731	5	0,120	1,177	-1,012
Item 666	1,264	0,364	3	0,372	0,841	-0,456
Item 667	1,539	2,698				
Item 668	0,475	1,253				
Item 669	1,176	0,031	3	0,272	0,695	-0,613
Item 670	0,813	1,179	2	0,814	0,415	0,992
Item 671	1,427	2,377				
Item 672	1,440	2,692				
Item 673	1,018	0,064	3	0,301	0,615	-0,456
Item 674	0,794	0,600	2	0,553	0,451	0,106
Item 675	0,460	-0,676	1	-0,505	0,334	-1,556
Item 676	1,345	1,124	4	0,555	0,742	0,282
Item 677	1,485	0,813	5	0,462	0,983	-0,119
Item 678	0,611	0,785	1	0,888	0,462	0,668
Item 679	1,081	1,912				
Item 680	2,132	-0,679				
Item 681	0,623	1,200	2	0,903	0,375	1,398
Item 682	0,984	0,265	2	0,337	0,562	-0,440

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 683	0,310	1,898	1	2,191	0,255	2,743
Item 684	1,535	0,053				
Item 685	1,292	0,257				
Item 686	1,897	-0,711				
Item 687	1,443	-0,655				
Item 688	1,668	-0,187				
Item 689	1,481	-0,626	5	0,134	0,898	-1,080
Item 690	1,471	-0,572	5	-0,422	0,888	-0,980
Item 691	1,174	1,254	5	-0,006	0,773	0,237
Item 692	0,669	0,626	2	-0,009	0,471	0,175
Item 693	1,722	0,321				
Item 694	0,762	0,746	3	-0,049	0,463	0,258
Item 695	0,377	1,255	1	0,892	0,228	2,823
Item 696	0,958	-0,532	3	-0,495	0,657	-1,183
Item 697	0,588	0,671				
Item 698	1,070	1,827				
Item 699	0,254	0,767	1	0,469	0,241	1,235
Item 700	1,206	0,539	4	-0,161	0,896	-0,047
Item 701	1,523	0,555				
Item 702	1,578	1,234				
Item 703	0,334	-0,298	1	-0,730	0,252	-1,524
Item 704	0,903	-0,217	3	-0,384	0,601	-0,846
Item 705	2,161	0,665				
Item 706	1,922	1,520				
Item 707	0,770	0,938	2	0,055	0,593	0,577
Item 708	0,600	-0,404	2	-0,520	0,465	-1,284
Item 709	0,894	1,289	4	0,064	0,651	0,629
Item 710	1,665	-0,011				
Item 711	0,273	0,616	1	0,315	0,238	1,329
Item 712	0,896	0,463	3	-0,135	0,604	-0,085
Item 713	0,381	1,305	1	0,951	0,237	2,406
Item 714	1,508	0,883				
Item 715	1,500	0,009	5	-0,295	0,933	-0,534
Item 716	0,928	0,490	3	-0,205	0,627	-0,302
Item 717	1,012	0,574	3	-0,110	0,652	-0,054
Item 718	1,664	0,921				
Item 719	1,292	2,776				
Item 720	1,274	-1,191	3	-0,716	0,815	-1,488

(Table A.1 continued)

	MIRT_A	MIRT_B	OPLM_A	OPLM_B	BILOG_A	BILOG_B
Item 721	0,561	0,962	2	0,112	0,349	1,096
Item 722	1,500	1,199				
Item 723	1,254	1,187				
Item 724	0,547	-0,865	3	-0,055	0,448	-1,313
Item 725	0,595	0,573	3	0,431	0,620	0,423
Item 726	0,377	0,232	2	0,333	0,399	0,011
Item 727	0,428	1,381	3	0,452	0,549	0,760
Item 728	0,707	2,502	4	0,619	0,804	2,084
Item 729	0,565	0,160				
Item 730	0,457	-0,504	3	-0,251	0,549	-1,446
Item 731	0,558	-0,222	3	-0,179	0,601	-1,040
Item 732	0,645	-0,271				
Item 733	0,563	-0,165	4	-0,097	0,617	-0,503
Item 734	0,635	1,600				
Item 735	0,896	1,572				

APPENDIX B

CORRELATION COEFFICIENTS OF ABILITY ESTIMATIONS WITH TURKISH PMS SUB-DOMAIN SCORES IN DIFFERENT SCENARIOS

Post-hoc simulations are conducted in various scenarios to determine the optimum algorithm. Correlation coefficients between PMS scores and Post-hoc simulation ability estimations are given in Table B.1, Table B.2 and Table B.3.

Table B.1 Correlation Coefficients of PMS Mathematics Assessments Sub-Domain Scores with ML and WML Ability Estimations of Post-hoc CAT Simulations under Different Starting and Fixed Test Reliability Termination Rules

		Termination rule: fixed test reliability		
Estimation	Starting rule	SE<0.20	SE<0.30	SE<0.40
ML	-1.50 < b < -0.50 (Easy)	0.65 with GE 0.68 with ME 0.72 with NU 0.59 with PS	0.63 with GE 0.63 with ME 0.66 with NU 0.52 with PS	0.55 with GE 0.56 with ME 0.61 with NU 0.47 with PS
	-0.50 < b < +0.50 (Moderate)	0.64 with GE 0.66 with ME 0.73 with NU 0.58 with PS	0.63 with GE 0.64 with ME 0.68 with NU 0.53 with PS	0.57 with GE 0.61 with ME 0.63 with NU 0.46 with PS
	+0.50 < b < +1.50 (Difficult)	0.65 with GE 0.67 with ME 0.73 with NU 0.58 with PS	0.61 with GE 0.64 with ME 0.66 with NU 0.49 with PS	0.60 with GE 0.62 with ME 0.65 with NU 0.51 with PS
WML	-1.50 < b < -0.50 (Easy)	0.65 with GE 0.65 with ME 0.71 with NU 0.54 with PS	0.63 with GE 0.65 with ME 0.68 with NU 0.54 with PS	0.57 with GE 0.61 with ME 0.65 with NU 0.52 with PS
	-0.50 < b < +0.50 (Moderate)	0.65 with GE 0.67 with ME 0.73 with NU 0.58 with PS	0.64 with GE 0.66 with ME 0.71 with NU 0.53 with PS	0.58 with GE 0.60 with ME 0.64 with NU 0.51 with PS
	+0.50 < b < +1.50 (Difficult)	0.65 with GE 0.68 with ME 0.72 with NU 0.59 with PS	0.61 with GE 0.65 with ME 0.68 with NU 0.52 with PS	0.57 with GE 0.59 with ME 0.62 with NU 0.47 with PS

**All correlations are significant at the 0.01 level.*

Table B.2 Correlation Coefficients of PMS Mathematics Assessments Sub-Domain Scores with ML and WML Ability Estimations of Post-hoc CAT Simulations under Different Starting and Fixed Test Length Termination Rules

		Termination rule: fixed test length		
Estimation	Starting rule	15 items	25 items	35 items
ML	-1.50 < b < -0.50 (Easy)	0.63 with GE 0.65 with ME 0.70 with NU 0.55 with PS	0.67 with GE 0.70 with ME 0.75 with NU 0.60 with PS	0.69 with GE 0.71 with ME 0.77 with NU 0.60 with PS
	-0.50 < b < +0.50 (Moderate)	0.65 with GE 0.67 with ME 0.71 with NU 0.56 with PS	0.67 with GE 0.70 with ME 0.74 with NU 0.59 with PS	0.69 with GE 0.72 with ME 0.76 with NU 0.60 with PS
	+0.50 < b < +1.50 (Difficult)	0.64 with GE 0.65 with ME 0.70 with NU 0.55 with PS	0.67 with GE 0.70 with ME 0.75 with NU 0.59 with PS	0.69 with GE 0.72 with ME 0.77 with NU 0.61 with PS
WML	-1.50 < b < -0.50 (Easy)	0.64 with GE 0.67 with ME 0.72 with NU 0.56 with PS	0.67 with GE 0.71 with ME 0.76 with NU 0.60 with PS	0.70 with GE 0.72 with ME 0.77 with NU 0.61 with PS
	-0.50 < b < +0.50 (Moderate)	0.64 with GE 0.68 with ME 0.72 with NU 0.56 with PS	0.67 with GE 0.70 with ME 0.74 with NU 0.60 with PS	0.69 with GE 0.73 with ME 0.77 with NU 0.60 with PS
	+0.50 < b < +1.50 (Difficult)	0.63 with GE 0.66 with ME 0.72 with NU 0.55 with PS	0.68 with GE 0.70 with ME 0.75 with NU 0.61 with PS	0.69 with GE 0.72 with ME 0.77 with NU 0.61 with PS

**All correlations are significant at the 0.01 level.*

Table B.3 Correlations of Ability Estimations between PMS Assessment and Post-Hoc CAT Simulation under the Use of Content and Exposure Control

	Average Number of Items Administered	Correlation with Weighted Mean
No content and exposure control	15.78	0.62 with GE 0.65 with ME 0.68 with NU 0.51 with PS
Content control only	12.77	0.62 with GE 0.61 with ME 0.66 with NU 0.53 with PS
Exposure control only	20.74	0.66 with GE 0.65 with ME 0.71 with NU 0.57 with PS
Both content and exposure control	17.09	0.64 with GE 0.66 with ME 0.72 with NU 0.61 with PS

**All correlations are significant at the 0.01 level.*

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name	: Gökçe, Semirhan
Nationality	: Turkish (TC)
Date and Place of Birth	: 27.02.1978, Dörtyol/Hatay
Marital Status	: Single
Phone	: +90 312 2104982
Fax	: +90 312 2104999
Email	: semirhan@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Secondary Sci. and Mathematics Education	2005
BS (Double)	METU Mathematics	2001
BS	METU Mathematics Education	2000
High School	Gaziantep Science High School	1995

WORK EXPERIENCE

Year	Institution	Enrollment
2000-Present	METU Medical Center (Computer and Automation Dept.)	Research assistant
2012 (March-July)	University of Twente (Res. Methodology, Measurement and Data Analysis Dept.)	Intern Student

FOREIGN LANGUAGES

English

HOBBIES

Computer technologies, brain teasers, math puzzles and bowling