EFFECTS OF DIFFERENT COMPUTERIZED ADAPTIVE TESTING STRATEGIES ON RECOVERY OF ABILITY

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

BY

İLKER KALENDER

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

MARCH 2011

Approval of the thesis:

EFFECTS OF DIFFERENT COMPUTERIZED ADAPTIVE TESTING STRATEGIES ON RECOVERY OF ABILITY

submitted by **İLKER KALENDER** 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 Math. Educ. Dept**

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

Examining Committee Members:

Prof. Dr. Fitnat Kaptan _____
Elemantary Education Dept., Hacettepe Univ.

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

Assoc. Prof. Dr. Nükhet Demirtaşlı _____
Educational Science Dept., Ankara University

Asist. Prof. Dr. Ali Eryılmaz _____
Secondary Science and Mathematics Educ. Dept., METU

Asist. Prof. Dr. Ömer Faruk Özdemir _____
Secondary Science and Mathematics Educ. Dept., METU

Date: *15. 03. 2011*

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 : İlker Kalender

Signature :

ABSTRACT

EFFECTS OF DIFFERENT COMPUTERIZED ADAPTIVE TESTING STRATEGIES ON RECOVERY OF ABILITY

Kalender, İlker

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

Supervisor : Prof. Dr. Giray Berberoğlu

March 2011, 140 pages

The purpose of the present study is to compare ability estimations obtained from computerized adaptive testing (CAT) procedure with the paper and pencil test administration results of Student Selection Examination (SSE) science subtest considering different ability estimation methods and test termination rules.

There are two phases in the present study. In the first phase, a post-hoc simulation was conducted to find out relationships between examinee ability levels estimated by CAT and paper and pencil test versions of the SSE. Maximum Likelihood Estimation and Expected A Posteriori were used as ability estimation method. Test termination rules were standard error threshold and fixed number of items. Second phase was actualized by implementing a CAT administration to a

group of examinees to investigate performance of CAT administration in an environment other than simulated administration.

Findings of post-hoc simulations indicated CAT could be implemented by using Expected A Posteriori estimation method with standard error threshold value of 0.30 or higher for SSE. Correlation between ability estimates obtained by CAT and real SSE was found to be 0.95. Mean of number of items given to examinees by CAT is 18.4. Correlation between live CAT and real SSE ability estimations was 0.74. Number of items used for CAT administration is approximately 50% of the items in paper and pencil SSE science subtest. Results indicated that CAT for SSE science subtest provided ability estimations with higher reliability with fewer items compared to paper and pencil format.

Keywords: Cat, ability estimation, test termination, science achievement, student selection procedure

ÖZ

FARKLI BİLGİSAYAR ORTAMINDA BİREYSELLEŞTİRİLMİŞ TEST STRATEJİLERİNİN YETENEK KESTİRİMİ ÜZERİNDEKİ ETKİLERİ

Kalender, İlker
Doktora, Orta Öğretim Fen ve Matematik Alanları Eğitimi Bölümü
Tez Yöneticisi: Prof. Dr. Giray Berberoğlu

Mart 2011, 140 sayfa

Bu çalışmanın amacı bilgisayar ortamında bireyselleştirilmiş (CAT) test yöntemi ile elde edilen yetenek kestirimlerini farklı yetenek kestirim ve test sonlandırma kurallarını dikkate alarak Öğrenci Seçme Sınavı (ÖSS) fen alt testinin kağıt kalem formatı sonuçları ile karşılaştırmaktır.

Çalışma iki aşamadan oluşmaktadır. İlk aşamada, ÖSS'nin CAT ve kağıt kalem formatlarından elde edilen yetenek kestirimlerini karşılaştırmak amacı ile post-hoc simulasyon uygulanmıştır. Yetenek kestirim yöntemleri olarak Maximum Likelihood Estimation ve Expected A Posteriori kullanılmıştır. Test sonlandırma kuralları ise standart hata eşik değeri ile sabit soru sayısıdır. Çalışmanın ikinci aşaması CAT performansını simulasyon dışında bir ortamda gözlemlemek amacı ile bir grup öğrenciye uygulanmıştır.

Post-hoc simulasyon bulguları CAT uygulamasının ÖSS için Expected A Posteriori yetenek kestirim yöntemi ile 0,30 ya da daha yüksek standart hata eşik değeri ile uygulanabileceğini göstermiştir. İki formattan elde edilen yetenek kestirimleri arasındaki korelasyon 0,95 olarak bulunmuştur. CAT ile kullanılan soru sayısı ortalaması ise 18,4 olmuştur. Gerçek bireylere uygulanan CAT ile kağıt kalem formatındaki ÖSS yetenek kestirimleri arasındaki korelasyon 0,74'tür. Gerçek bireylere uygulanan CAT ile bireylere sorulan soru sayısında yaklaşık %50 oranında düşüş sağlanmıştır. Sonuçlar CAT formatının ÖSS fen alt testi için kağıt kalem testi ile karşılaştırıldığında daha yüksek güvenilirliğe sahip yetenek kestirimlerini daha az soru ile sağladığı göstermiştir.

Anahtar Kelimeler: Cat, yetenek kestirimi, test sonlandırma, fen başarısı, öğrenci seçme yöntemi

To Özge

ACKNOWLEDGEMENTS

I wish to express my gratitude to Prof. Dr. Giray Berberoğlu for his sincere interest, patience and criticism throughout my study.

Also I am thankful to members of examining committee for their comments and suggestions.

I wish to thank my wife for sharing all the stress with me and encouraging me.

I would to thank administrators, instructors and students of the Preparatory School of Middle East Technical University for their valuable contribution to my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

1PL	One-parameter Model
2PL	Two-parameter Model
3PL	Three-parameter Model
CAT	Computerized Adaptive Testing
CTT	Classical Test Theory
EAP	Expected A Posteriori
ICC	Item Characteristic Curve
IIF	Item Information Function
IRT	Item Response Theory
MAP	Maximum a Posteriori
MLE	Maximum Likelihood Estimation
P&P	Paper and pencil test
SE	Standard Error
SSE	Student Selection Examination

CHAPTER 1

INTRODUCTION

Paper and pencil tests have been a dominating method for measuring individuals' ability levels for many years. This method includes giving examinees a fixed number item in a fixed order (and generally in the same order) and examinees tend to follow the order of items in the booklets. This format provides a highly-standardized measurement methodology (Weiss, 1983).

Large-scale test administrations in Turkey are conducted in a way that all examinees take a paper and pencil test including the same items at a certain date over Turkey. The Student Placement Examination, the Foreign Language Examination for Civil Servants, the Entrance Examination for Graduate Studies, etc. are some of the tests that are taken by many examinees. For example, over one million examinees take the Student Selection Examination (SSE) each year. The Foreign Language Examination for Civil Servants is also taken by thousands of examines (Student Selection and Placement Center, 2010).

Among those exams, SSE that is conducted once a year, has a special importance since scores obtained from SSE are used for selection and placement of students to higher education programs in Turkey. One of the principle criticisms on SSE is the fact that difficulty of the items and ability levels of the students do not match. This is a fact that can be confirmed by checking the means of subtests. For example, means of science subtests for the year of 2009 and 2008 are 4.0 and 3.9 out of 30, respectively. Low mean scores for SSE science subtest indicate that there is a problem in balancing item difficulties and examinees' ability levels. Giving unmatched tests to examinees in terms of their ability levels may produce unhealthy item and test parameters and unreliable test scores.

In unmatching paper and pencil tests, since each examinee is given the same items, it is highly possible to receive items that are too easy or too hard. Items that are not suitable for examinee's level provide little information about his/her ability level. Therefore, many items are needed to obtain reliable estimations of abilities. Also, giving inappropriate items in difficulty to examinees can make them bored, tired, etc. and can be waste of time. In addition, examinee can be doing blind guessing due to items that are difficult and this, in turn, increases error of ability estimation. If it is possible to give each examinee a test with an ideal matching to his-her ability level, the problems mentioned above could be solved effectively (Mead & Drasgow, 1993).

But in this situation, a paradox seems to arise: if an examiner knows the examinee's ability level, then there is no need to test him/her. On the other hand, if ability level is unknown, how to structure a test tailored to examinee's ability level. As a solution to this situation that can be called paradox of test design, it has been suggested that to determine the examinee's ability level, responses previously given in the test could be used to select the next appropriate items for an examinee (Weiss, 1983).

1.1 Computerized Adaptive Testing

In tailored tests (or adaptive tests suggested by Lord [1980]), items are dynamically selected after each response using the responses given by examinees and next items are selected that is best match to examinee's ability levels from an item bank. Therefore examinees receive items that are most appropriate to their ability levels. Even tough this approach seems fine theoretically, practical applications can be limited due to speed and time requirements. On other hand, using computers for adaptive tests can provide a solution.

Before come to the term computerized adaptive testing (CAT), historical development of adaptive testing starts with the idea adaptive testing.

Adaptive testing idea first arose by Alfred Binet's IQ Test (Binet & Simon, 1905). This test constitutes a first example for an adaptive test with its all

essential features. Friedrich Lord from Educational Testing Service made significant contributions to adaptive testing literature. It is interesting to observe that a significant proportion of developmental ideas about adaptive testing came from studies of American Navy (Weiss, 1983). With the advancement in computer technology, concept of adaptive testing had a transformation to the idea computerized adaptive testing.

Basic idea behind the computerized adaptive testing is to give examinees items only tailored or adapted to their ability levels. By this way, several advantages can be obtained such as a significant reduction in the number of items given. Among advantages of CAT over conventional testing, Betz and Weiss (1974) state that they (i) are shorter and (ii) provide reliable ability estimates of examinees.

Embretson (1996) states that CAT administration needs fewer items, producing more valid measurement experiences than paper and pencil tests which includes more items. Also Rudman (1987) calls the CAT as the measurement method of 21st century.

Advantages and disadvantages of Computer Adaptive Tests can be listed as follows (Cikrikci-Demirtasli, 1999; Hambleton & Swaminathan, 1984; Lord & Stocking, 1968; Rudner, 1998; Sands, Waters & McBride, 1997):

- Time required for implementation of test become lesser,
- Each examinee receives a test tailor to his/her ability level,
- Security is increased since printed question sheet is not used therefore transportation
- Scoring can be made immediately after testing,
- Test can be given any time,
- Need for paper and pencil diminishes,
- Update of item pool and inclusion or exclusion of items are easy,
- Test standardization is achieved,
- Flexibility regarding item selection is increased,

- Item formats that are not possible to deliver in paper and pencil tests can be used including multimedia, animation, user interaction, etc.

Beside the advantages, CAT applications have some disadvantages:

- Need for use computer in testing session make people with computer anxiety feel uncomfortable,
- Computer hardware limitations and cost can be a problem,
- Failure to meet the criteria of unidimensionality of trait measured (Unidimensionality means that a measured trait has a single factor affecting it),
- Need for a large item pool.

Computer anxiety can especially be regarded important because of its relatedness to human characteristics. However, Legg and Buh (1992) reported no significant differences between attitudes and anxiety levels towards CAT for subgroups including different socio-cultural levels.

Also there are some potential problems that can be arisen when working with CAT:

- Only one item is displayed at a given time,
- To skip a given item it is required to provide a response,
- Moving among items is not allowed unless a response is provided.

1.2 CAT Administrations

There are several large scale testing programs including CAT administration.

GRE (Graduate Record Examination) is an examination, results of which are used for admission to graduate schools in USA. GRE was developed and conducted by Educational Testing Service (a.k.a. ETS). A similar examination to

GRE is GMAT (The Graduate Management Admission Test). It is a standardized test including mathematics and English language items and was developed under supervision of Graduate Management Admission Council (GMAC) to be used by business schools. GMAT is mainly given in CAT format where it is possible to. Also TOEFL (Test of English as a Foreign Language), a test for measurement English language proficiency levels of non-native English speakers, have been given in adaptive format through its history (GMAT, 2010; GRE, 2010; TOEFL, 2010).

CAT programs are also used for achievement testing. For example, Papanastasiou (2003) stated that CAT as the most efficient and advantageous computer-based measurement experience for science. Measures of Academic Progress (MAP) (Northwest Evaluation Association [NWEA], 2010) for Science including dimensions of concepts, processes and general science for primary level students Mathematics Assessment for Learning and Teaching (MALT, 2010) is given 5 to 14-aged students for Mathematics assessment diagnosis. Another test is Scholastic Math Inventory (SMI, 2010) used to determine skills of mathematics of examinees. These are all CAT administered measurement programs.

For Turkey, a country in which large-scale testing programs are conducted widely, CAT administration can be a potential solution to the problems associated with using paper and pencil tests.

SSE includes one qualitative and one quantitative subsection. In qualitative part there are Turkish, history, geography and philosophy subtests, while quantitative part is constituted from mathematics, physics, chemistry, and biology subtests. Items of SSE are related to reading comprehension in Turkish language, and thinking abilities using basic concepts and principles in mathematics and science. Though items are developed based on curriculum, SSE mainly assesses higher-order thinking abilities covered in courses in high schools (Student Selection and Placement Center, 2010),

Using CAT administration for SSE can provide a number of advantages. First, each examinee who takes SSE receives a test that matches to his or her

ability level. By this way, item parameters of good quality and reliable test scores would be obtained. In paper and pencil format of SSE, items do not match to examinees' ability levels and item parameters are of low quality. Providing a correct response to any item could change examinees' ordering significantly. If an examinee gives a correct response to an item by blind guessing he/she could receive higher scores than he/she deserves due to poor item parameters. CAT gives each examinee a test that is tailored to their ability levels and examinee take appropriate items in difficulty. Therefore, examinees exhibit aberrant testing behaviors less such as cheating, anxiety arising from difficulty of test, etc. Also limitations of paper and pencil format in terms of item formats could be overcome by CAT administration. New item formats such as interactive items, multimedia items, etc. can effectively be used in CAT administration. Test scoring is made instantly and therefore there is no need to answer sheeting reading process. Also issues of security and transportation of test booklets are eliminated. Copying detection is also conducted easily by CAT administration since computer records a lot of data for each examinee. There are copying or collusion detection methods applied for CAT administrations (van der Linden, 2008; Wise & Kong, 2005).

1.3 CAT Studies

Koklu (1990) made a comparision between adaptive and paper and pencil formats with respect to validity and reliability. Koklu reported no statistically significant difference between reliability estimations of adaptive and conventional format. On the other hand, Koklu stated that although differences were not high, adaptive administration provided better results.

Kaptan (1993) compared ability estimates obtained from paper and pencil test and computer adaptive test. In her study, a test was formed using the mathematics items and examinees received a 50-item paper and pencil test and 14-item computerized adaptive test. Ability estimation was conducted using maximum likelihood estimation method by a computer program developed by the researcher and results indicated a 70% reduction rate in the items administered by

CAT format. Also no significant difference was reported between two methods in ability estimations by the researcher.

There are several dimensions of a CAT administration that affects outputs. Among them are item selection, item exposure control, and, ability estimation and test-stopping rule.

Item selection procedure means the methods to select next item to give examinees. Item exposure control includes approaches balancing proportions of items from different subdomains to keep content validity.

Ability estimation methods include approaches for estimating examinees' abilities. There are four approaches exist for ability estimation in the literature: (i) Maximum Likelihood Estimation (MLE) (Birnbaum, 1958) and, (ii) Owen's Bayesian Estimation (OWEN) (Owen, 1969), (iii) Expected a Posteriori Estimation (EAP) (Bock & Aitken, 1981), and (iv) Maximum a Posteriori estimation (MAP) (Samejima, 1969). Among them MLE and EAP gained popularity.

MLE estimates the ability by using joint probability and then finding the point that maximizes the ability estimation.

Likelihood function of ability is defined as follows:

$$L(u \mid \theta) = \prod_{i=1}^{n} P_i i(u_i \mid \theta)$$

where u_i response to item i in a response vector u, P is probability of true response for item i

Maximum likelihood of ability, θ, is expressed by $\hat{\theta}$ which is a value that maximizes the likelihood function.

$$\ln L(u \mid \theta) = \sum_{i}^{n} \ln P_i(u_i \mid \theta) = \sum_{i}^{n} \ln P_i^{u_i} Q_i^{1-u_i}$$

where Q_i is equal to 1-P

Taking first partial derivative of the likelihood function and setting it equal to zero, maximum value of θ can be found.

$$\frac{\partial \ln L(u \mid \theta)}{\partial \theta} = \sum_i \frac{P_i'(u_i \mid \theta)}{P_i(u_i \mid \theta)} = \sum_i \frac{(u_i - P_i)P_i'}{P_i Q i} = 0$$

MLE has a computationally-easy formula and it is also unbiased, has no a priori definitions, but produces higher standard errors (Hambleton & Swaminathan, 1984). On the other hand, MLE has a requirement that examinees provide one correct and one wrong response to obtain a maximum point to make estimation. Without that, MLE cannot produce ability estimations.

Bayesian EAP method computes posterior distribution for an examinee's ability using prior distribution and it uses mean of the ability distribution unlike MAP which uses mode of the ability distribution.

$$p(u \mid \theta) = \frac{L(u \mid \theta)g(\theta)}{P(u)} = \frac{L(u \mid \theta)g(\theta)}{\int L(u \mid \theta)g(\theta)d\theta}$$

where g(θ) is the prior information about examinees' abilities.

$$E(\theta \mid u) = \int_{-\infty}^{\infty} \theta p(\theta \mid u) d\theta$$

and

$$Var(\theta \mid u) = \int_{-\infty}^{\infty} \theta^2 p(\theta \mid u) d\theta - (E(\theta \mid u))^2$$

where E(θ|u) and Var(θ|u) are the mean and variance of the posterior distribution.

To solve these integrals, approximations proposed by Stroud and Sechrest (1966) can be used:

$$\hat{\theta} \equiv E(\theta \mid u) = \frac{\sum_{k=1}^{q} X_k L(X_k) W(X_k)}{\sum_{k=1}^{q} X_k L(X_k)}$$

and

$$\hat{\sigma}^2(\hat{\theta}) = Var(\theta \mid u) = \frac{\sum_{k=1}^{q} (X_k - \hat{\theta})^2 L(X_k) W(X_k)}{\sum_{k=1}^{q} L(X_k) W(X_k)}$$

where X_k is one the quadrature points. $W(X_k)$ is a factors related to quadrature point. And $L(X_k)$ is conditioned likelihood function at quadrature points.

Test-stopping rule is another dimension studied in CAT literature. To end a CAT session, there are several criteria stated in the literature (Simms & Clark, 2005): (i) fixed number of items (De Ayala, 1992) and (ii) standard error threshold, (iii) information of an item below a predefined value, (iv) combined use of the preceding rules. Of these rules, fixed number of items and standard error threshold are widely used methods in CAT administrations (Gushta, 2003; Weiss, 1983). Using standard error threshold is objected by some arguing this rule is biased. (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001). On the other hand, Babcock and Weiss (2009) found that this rule is no biased than rules based on administering fixed number items to examinees.

Fixed number of items ends a session after a predetermined number of items is given to examinees. This approach favors the content validity since it is certain that examinees a certain number of items. However, reliability of the test is not guaranteed in this approach. After all items are given to examinees, standard error may still be too high to be reliable. To solve this, a standard error threshold can be defined prior to CAT administration. By this way, it is assured that ability estimations for all examinees are reliable. But this approach may violate content validity since test can be ended without items from some subdomains are given. Or more seriously, due to students who provide aberrant response patterns (correct responses for very hard items for an examinee with low ability, or vice versa) standard error never reaches to level defined a priori. For a review of test stopping rules, Hambleton, Zaal and Pieters (1991) can be investigated.

1.4 Definition of Terms

CAT: A testing methodology to give examinees tailored or adapted tests to their ability levels.

SE: Standard error of ability estimation. In CAT, SE is used as a reliability measure and test termination value.

Information: Information is defined in the form of a function of ability and item parameters. It is related to SE by $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$. Information is used to select next items from item bank and calculate SE.

Classical Reliability: Classical test theory uses several reliability coefficients to indicate reliability such as Pearson's correlation coefficient. On the other hand, in IRT, in turn, CAT reliability is given by SE. Conversion between reliability coefficients of CTT and IRT is conducted by using the formula:

$$SE = \sqrt{1 - r}$$

1.5 Purpose of the Study

The purpose of the study is to compare CAT and paper and pencil test results in SSE science subtest through a simulation and a real study by considering different ability estimation methods (MLE and EAP) and termination rules (fixed number of items vs. fixed SE test).

First using post-hoc simulation techniques based on responses of examinees to past Student Selection Examinations, ability estimations were obtained and these estimations were compared to examinees' scores from paper and pencil SSE science subtest. Then using live individuals a real CAT application is conducted using an item bank including past science items of SSE so that ability estimations of real students from paper and pencil test and CAT administration were compared. Live testing phase is especially important for the present study. In post-hoc simulation phase responses of examinees given to paper and pencil format of SSE science subtest are used. Those responses are not provided especially for CAT administration in front of a computer by examinees, they are responses given for P&P SSE science subtest. Therefore, there is no effect arising from CAT administration on examinees. On the other hand, by conducting a live CAT administration may provide more realistic picture of ability estimates. Based on that reason, a live CAT testing phase is included to the present study. Another reason for live CAT is that it uses a large item bank and provides a realistic administration. In post-hoc simulation CAT is simulated with an item bank including the same number of items with P&P SSE science subtest. Thus findings of the simulation phase are limited and that limitations can be overcome by conducting a real CAT.

For the present study, two ability estimation methods (MLE and EAP) methods are investigated because of easy computability and lower standard errors.

In addition to ability estimation methods, two test stopping rules (fixed SE and fixed test length) are also included to the present study. Comparisons among ability estimations from CAT and paper and pencil are made on samples from different high schools (state, Anatolian, and private) and using different test lengths to observe performance of CAT administration on different test and examinee groups.

Very low means for total scores of SSE science subtest led the research to investigate the applicability of CAT format for that subtest. Low mean due to unmatched items with ability levels of examinees provide unreliable test scores and item parameters of low quality. Also the fact that missing rates of science subtest are too high is another factor that directed the researcher to investigate CAT format of SSE science subtest.

Based on that, research problem for the present study can be stated as follows:

1. Does CAT administration of SSE science subtest estimate ability levels of examinees compared to paper and pencil format for different school types and different test lengths?
    1.1. Do post-hoc simulations provide reliable and comparable ability estimates to paper and pencil format?
    2.1. Does live CAT administration provide reliable and comparable ability estimates to paper and pencil format?
2. Do different ability estimation methods (MLE vs. EAP) produce differences in ability estimation?
3. Do different test termination rules (fixed number of items vs. fixed SE) produce differences in ability estimation?

1.6 Significance of the Study

Large-scale testing administrations are widely used in Turkey (Student Selection Examination, Graduate Education Entrance Examination, etc). It is known that paper and pencil format of large-scale testing programs have many problems. Investigation of CAT format for SSE science subtest makes a significant contribution for findings alternatives of test administration formats to select students for transition to higher education.

Since the present study includes comparisons of ability estimation and test termination rules, outputs of different CAT testing strategies can be compared for SSE. Using different school types representing different ability groups also provide important results in applicability for CAT to sub groups of SSE takers.

The present study (i) makes a contribution to studies seeking alternatives for methods of selecting and placing student to higher education programs from the dimension of measurement techniques and (ii) provides insight for people those who are related to educational sciences and to educational policies about an alternative test format for SSE.

Moreover, some issues regarding the explanation of CAT to the public are discussed in the present study. Because of its nature, for example each examinee will be given items differing by difficulty, content, etc. and these are likely to arise public concern on the reliability of the examinations from the view of those related to these examinations such as examines, families, etc. These points are required to be well explained.

Though only science subtest is at focus for the present study, it is expected that results that can be obtained could be generalized to any large-scale testing administered by Student Selection and Placement Center in Turkey through the similar analyses.

CHAPTER 2

LITERATURE REVIEW

In this chapter, literature related to CAT procedures is presented.

Research dimensions of computer adaptive test can be grouped by several ways. One of the categorization was made by Weiss (2010) and includes the following dimensions: (i) Content Balancing**,** (ii) Estimation Method, (iii) Stopping Rule, (iv) Multiple Scales, (v) Right to Go Back Responded Items, (vi) The Worldwide Web and CAT, etc.

Since the present study deals with ability estimation methods and test stopping rules of the dimensions stated above, rest of the dimensions will be out of the focus.

Weiss (2010) also stated research approaches that are used for CAT. These are live-testing studies and two kinds of simulations: real-data or post hoc simulations, and Monte Carlo simulations.

Live testing involves implementation of real test to real examinees (Weiss, 1983). Real-data or post hoc simulations are conducted in order to determine how number of items in a test could be reduced without any loss in psychometric properties of the test scores and uses real responses of live examinees to paper and pencil tests. Monte Carlo studies are used to evaluation of performances of different computer adaptive testing applications with real or simulated data sets (Harwell, Stone, Hsu & Kirisci, 1996).

CAT administration is widely used around the world. Economides and Roupas (2007) evaluates CAT systems: Graduate Management Admission, Test (GMAT), Graduate Record Examination, (GRE), Test of English as a Foreign, Language (TOEFL), Microsoft Certified, Systems Engineer (MCSE), Cisco, the

Computing Technology Industry Association (CompTIA), etc. Researchers stated that factors such as security, reliability have priority over giving examinees feedback to examinees and they provide some suggestions for feedback provided to examinees.

CAT administrations are used in the followings (Weiss, 2010):

- GMAT (Graduate Management Admission Test) (http://www.mba.com/mba/TakettheGMAT/TheEssentials/WhatIstheGMAT/ComputerAdaptiveFormatNEW.htm),
- GRE (Graduate Record Examination),
- CITO (http://www.cito.com/en.aspx)

Also,

- Adaptive Matrices Test (AMT),
- ASCP (American Society of Clinical Pathologists-Board of Registry Certification Examinations),
- ASVAB (The Armed Services Vocational Aptitude Test Battery),
- CAT of Written English for Spanish Speakers,
- BULATS (Business Language Testing Service) Computer Test,
- CATE (Computerized Adaptive Test of English),
- COMPASS series of tests from ACT,
- LPCAT (Learning Potential CAT),
- MAP (Measures of Academic Progress),
- Microsoft Certified Professional Examination,
- NAPLEX (North American Pharmacist Licensure Examination),
- NCLEX (National Council Licensure Examinations),
- STAR Math, Reading, and Early Literacy.

2.1 CAT Administrations

First implementation of tailored, or adaptive test, was used by Alfred Binet as IQ Test (Binet & Simon, 1905). Binet's work includes all the characteristics

that an adaptive test application is expected to have: a preset item pool, items grouped with respect to difficulty levels, starting choice, a predefined scoring methodology, a selection rule for items to be drawn the item pool and predefined a stopping rule. Even tough this first implementation of CAT seems very simple; it provides a basis for further applications.

Later on 1950s except some studies there were no progress in the field. In 1960s, Friedrich Lord from Educational Testing Service made significant contribution to the field. Lord's main idea was that (1980) "a fixed-number-item test is not appropriate for examinees with higher and lower ability levels. If items tailored for the examinee's ability level were used, testing could be done without any loss of information. Then the field continued to develop by studies conducted by American Navy (Weiss, 1983).

Weiss and Betz (1973) made a review of research about adaptive ability testing of the time starting from Binet's work (1905) that is considered to be the first about adaptive measurement. The researchers discussed strengths and weaks of adaptive measurement in detail and also states potentials and problems related to that new measurement approach. They listed the advantages of adaptive testing administration as follows: fewer items than conventional testing, higher reliability, and more valid test. At the end of the study, they also pointed to potential problems that could arise with using adaptive testing. Researchers favored use of adaptive testing and with increasing availability of computer, they pointed to computerized adaptive testing.

Since synchronized complex calculations, quick drawing of items from the item bank and selection the next items based on information functions are needed, it was not until 1970s that notion of Computer Adaptive Test (CAT) arose on that years affordable computers with higher capabilities became available (Cikrikci-Demirtasli, 1999).

Betz and Weiss (1974), in a another study, used a Monte-Carlo simulation to assess psychometric properties of an adaptive testing and compare ability estimates obtained from adaptive and paper and pencil administration.

Researchers reported that adaptive testing administrations yielded higher reliability than conventional testing administration.

Mills and Stocking (1996) discussed practical dimensions of CAT such as starting item, ability estimation method, test stopping rules, etc. This research excellent provided all basic issues from a practitioner perspective.

Mead and Drasgow (1993) conducted a meta-analysis study to investigate equivalence of CAT and paper and pencil format. Researchers examined 159 correlations (123 speed tests and 26 power tests). For speeded test administration correlation was found to 0.91 and for power test 0.72. Combined correlation without splitting administration type was 0.91. Based on these findings, researcher reported no differences in equivalence between according to test administration format.

McBride and Martin (1983) compared computer adaptive test and paper and pencil tests in terms of validity and reliability and stated that with computer adaptive tests highly reliably results were obtained using only half of the items in paper and pencil test. Also calibrations with larger samples give better and more valid results. As the result of their studies, a 15-item computer adaptive test yielded comparable results with paper and pencil tests with the same length.

Engelhart (1986) conducted a Monte-Carlo simulation to find out the effects of misspecified IRT model to CAT administration. To this end, researcher produced a virtual item bank and responses of examinees to those items. Items were generated to fit 2-Parameter model however researcher used Rasch model to obtain a misfit situation. Researcher stated that using misfit model for CAT administration produces biased ability estimations but this could be minimized using a modified ability estimation method. Researcher also reported increase in the number of items administrated in CAT sessions did not a make a significant contribution to eliminate or minimize the biased ability estimation.

Ben-Porath, Slutske and Butcher (1989) conducted a real data simulation to find reduction rate of items administered in CAT session compared to paper and pencil format of Minnesota Multiphasic Personality Inventory (MMPI). In

their study, researchers used responses of people given to paper and pencil format of MMPI for simulated CAT administration. Researchers used different testing strategies to observe their effects in CAT format such as cluster item administration. At the end of the study, researchers reported significant reduction rates compared to paper and pencil format of MMPI.

2.2 Achievement Testing by CAT

Koklu (1990) compared adaptive and paper and pencil format of a test with respect to validity and reliability. Koklu reported no statistically significant difference between reliability estimations of adaptive and conventional format. On the other hand, the researcher investigated the relationship between test scores from adaptive and paper and pencil formats, and grades of participants' science courses to investigate validity of testing formats and found correlation coefficients of 0.88 and 0.81 for adaptive and conventional testing format respectively. Koklu stated that although differences were not high, adaptive administration provided better results.

Kaptan (1993) made a comparison between ability estimates obtained between paper and pencil tests and computer adaptive tests. In her study, a test was formed using mathematics items. Examinees received a 50-item paper and pencil test and 14-item computer adaptive. Ability estimation was conducted using MLE by a computer program developed by the researcher and results indicated a 70% reduction rate in the items administered by CAT format. Also no significant difference was found between two methods in ability estimations by the researcher.

Cikrikci-Demirtasli (1999) introduces systematic of CAT comprehensively. In her study, researcher introduces CAT principles, different CAT forms, etc. and discusses usability of CAT format.

In his dissertation, Iseri (2002) used an item pool including items from Secondary School Student Selection and Placement Examination. He stated that computer adaptive tests estimated students' achievement levels using fewer items.

In test sessions in which students were allowed to go back to the items responded earlier, estimations for students with higher ability level was better than those with lower levels. Bayesian estimation method made better ability estimation and both stopping rule using a fixed number of items and with fixed error of measurement yielded reliable results.

Miller (2003), in his dissertation, compared CAT and conventional testing for achievement levels of students competencies that states in USA defined. Researcher gave 267 students both paper and pencil and CAT formats of the same test and found that scores estimated by CAT are significant correlations with P&P scores. Miller also asked students to identify their preferences on test format. Results indicated no significant differences between students' preferences for test format.

In a dissertation by Yasar (1999) KR-20 reliability coefficients of CAT were investigated. Researcher compared correlations obtained from CAT and paper and pencil format of the same test. In the study CAT item bank includes only 61 items. Correlation between two different formats was found significant with a coefficient of 0.36, indicating a low relationship. Researcher indicated some potential reasons for that such as limited number of items in the bank, and test stopping rule with fixed number of items.

Eggen and Straetmans (2000) conducted a study in which CAT is used for classification of examinees into one of three groups in Netherlands. The purpose of the study was to compare quality of CAT administration of a placement test used for student to courses according to their ability levels. At the end of the study, researchers reported a 22% to 44% reduction in the number of items required for CAT compared to paper and pencil test.

2.3 Item Response Theory

Since adaptive testing, regardless of computerized or not, depends mainly on items, another testing framework rather than Classical Test Theory (CTT) is needed. Item parameters such as item difficulty, item discrimination, etc. on CTT

are estimated from groups who responded to a group of items and their values are dependent on that group. This makes constructing test difficult for groups of individuals with ability levels who have different than original group. Also examinees' ability level estimations are also test-dependent which make comparisons very hard among individuals who take different tests with different items.

So Classical Test Theory is not so healthy for adaptive testing that needs group independent item characteristics, item-independent ability estimations, and individual reliability estimations.

Mathematical theory used in computer adaptive testing applications is Item Response Theory (IRT) that address all of the points stated above. (Embretson & Reise, 2000, De Ayala, 2009). IRT provides a standard framework for estimating ability levels of individuals (Hambleton; Swaminathan & Rogers, 1991).

Two postulates that IRT arise from are (i) examinee performances can be estimated by some latent traits (IRT can also be called Latent Trait Theory), and (ii) relationship between performances of examinees and latent traits can be depicted by item characteristic functions or item characteristic curves that are monotonically increasing functions.

The most striking features of the IRT is that (i) it gives ability estimations independent of items used in the tests, meaning that if the same examinee are given two different sets of item items, ability levels estimated fort this examinee do not differ. And (ii) item parameters are estimated independently of population, which means that items parameters would be the same regardless of the calibration group due to nonlinear regression techniques of IRT. These two features are called invariance of ability parameters and invariance of item parameters, respectively. Another unique feature of IRT is that it provides individual standard error estimates that are individual reliabilities. (Embretson & Reise, 2000)

IRT has two main assumptions that are required to be met for using IRT models. *Unidimensionality* means that only one psychological trait is measured by

the items that make up the test. This assumption cannot be strictly met due to nature of trait measured. For multidimensional traits, there are several models such as Multicomponent Latent Trait Model developed by Whitely (1980). *Local independence* means that when the abilities affecting test performance are held constant, examinees' responses to any pair of items are statistically independent. That is, no relationship exists between examinees' responses to different items. When unidimensionality assumption is met, local independence is also regarded met. But that is not true for the opposite. For other non-dichomotic IRT models, see Van der Linden & Hambleton (1996).

There are several models proposed in IRT: these models include, one parameter (1PL Model), two parameters (2PL Model), and three parameters (3PL Model). 1PL model has a only single parameter named item difficulty, *b*. 2PL adds an additional parameter to *b*, item discrimination, *a*. Finally 3PL has an extra parameter, *c*, pseudo-guessing parameter.

The IRT models are defined in two forms: in normal ogive functions and in logistic functions. All dichotomous IRT models can be seen below (Hambleton & Swaminathan, 1984).

One Parameter Logistic Model:

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \quad \textit{where i = 1, 2, 3...n}$$

Two Parameter Logistic Model:

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \quad \textit{where i = 1, 2, 3...n}$$

Three Parameter Logistic Model:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \quad \textit{where i = 1, 2, 3...n}$$

One Parameter Normal Ogive Model:

$$P_i(\theta) = \int_{-\infty}^{\theta - b_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Two Parameter Normal Ogive Model:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Three Parameter Normal Ogive Model:

$$P_i(\theta) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

In addition to the three model given above, there is a model including four parameter proposed by both in logistic and normal ogive forms. In that model similar to c parameter which defined a chance level for examinees with the lowest ability, a parameter represented by γ is included to define a probability that an examinee from the highest ability levels provide false response by chance.

Four Parameter Logistic Model:

$$P_i(\theta) = c_i + (\gamma_i - c_i) \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \quad \text{where i = 1, 2, 3…n}$$

Four Parameter Normal Ogive Model:

$$P_i(\theta) = c_i + (\gamma_i - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

1PL model is also known as Rasch model (Rasch, 1960), even tough original model proposed by Rasch do not resemble to 1PL logistic or ogive

models. It has structural similarity to 1PL normal and ogive models in using only one item parameter, item difficulty, and producing the same curve with 1PL logistic model. Model proposed by Rasch is as follows.

$$\xi_{ij} = \frac{\delta_t}{\theta_j}$$

and

$$P_i(U_{ij} = 1 \mid \theta) = \frac{\theta_j}{\theta_j + \delta_j}$$

For 1PL and 2PL parameter models, *b* parameter is the point ability on scale for which probability for correct answer is 0.5. For all models *a* parameter is proportional to slope of the curve at the point $\theta = b_i$ and indicates item discrimination. Higher slopes mean items with higher discrimination. And last parameter, *c*, is called pseudo-guessing parameter and represents the probability that an individual provide a correct response by chance without ability to give correct response. Higher the values of c are, greater the chance individual provide correct response (Rudner, 1998). For 3PL b is equal to $(1 + c_i)/2$ since there is a chance factor that increases minimum probability of correct response from zero to upper levels.

In literature logistic models are usually more used since they are mathematically easier to compute. Normal ogive models include integration, while logistic models are just parametric functions. A scaling factor, D, is used to approximate logistic models to normal ogive models. Haley (1952) indicated that when D is equal to 1.7, difference between normal ogive and logistic models is minimized.

Among the models, 1PL was proved to perform better for small samples and to be robust to violation of assumption of IRT (Lunz & Bergstrom, 1994), on the other hand, it characterizes examinee behavior using only one parameter, item discrimination (*a*).

Probability that an examinee give a correct response to an item calculated based on these item parameters versus ability level constitutes a graph named Item Characteristic Curve (ICC). Figure 2.1 below shows four different IRC with different item parameters.

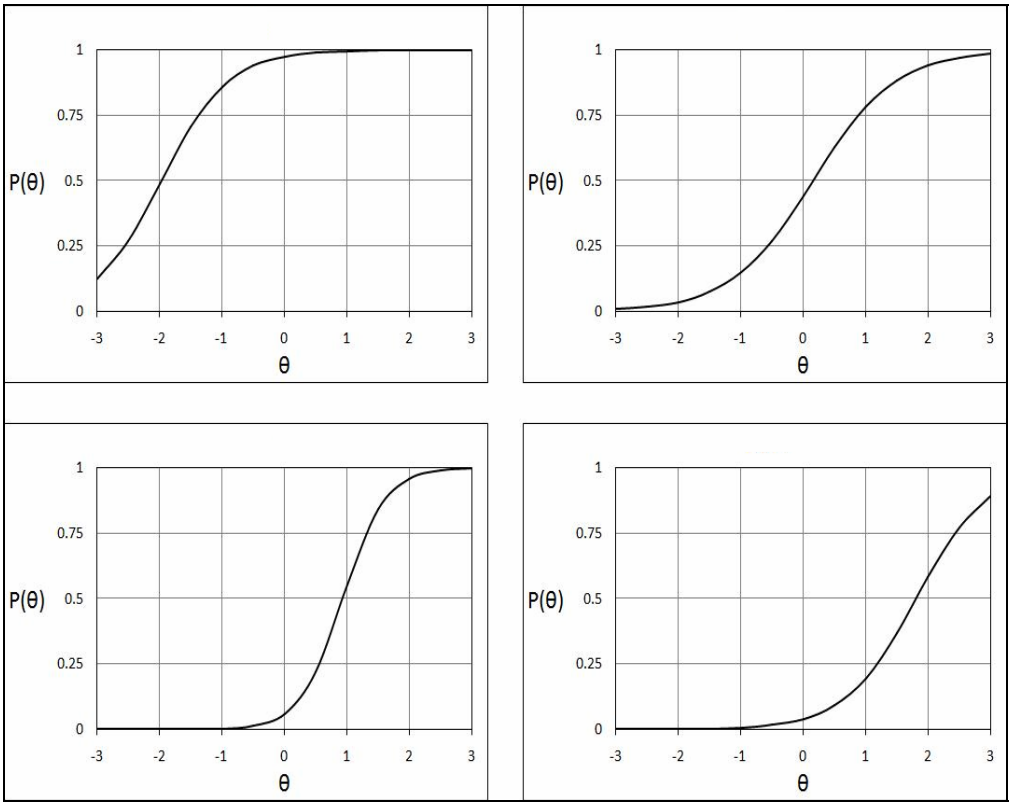

Figure 2.1 Different ICCs

In the graphs above, x-axis represents ability levels and y-axis represent probability that an examinee give a true response to that item, a parameter gives x-value of inflection point at 0.50 probability and make the graph steeper or flatter. b move the graph left (easier item) and right (harder item). An idealized characteristic curve is shown in Figure 2.2. It well discriminates between low and high ability students. But unlike theory, it is impossible to obtain such prefect curves practically.

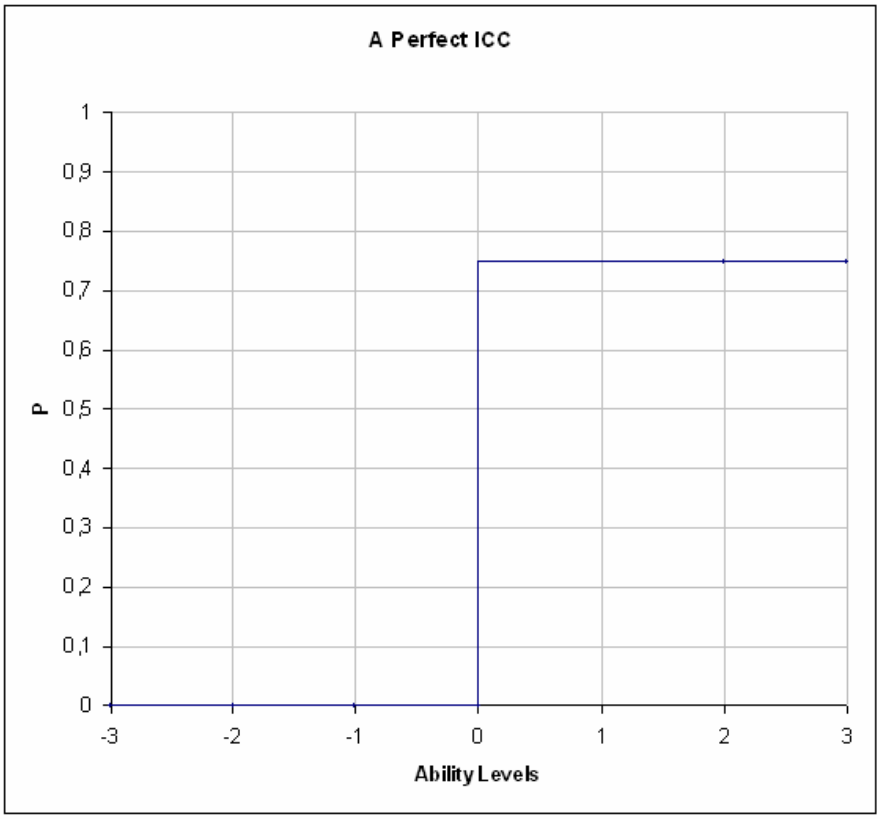

Figure 2.2 An idealized Item Characteristic Curve

Another important feature of IRT is the Item Information Function (IIF). In CAT sessions, IIFs are used for (i) calculated SEs for finding test reliability and (ii) selecting next item to be given to examinees. In the present study, item selection is based on selecting items with the highest item information at the ability levels of examinees (Rudner, 1998).

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{[P_i'(\theta)]^2[Q_i'(\theta)]^2} \text{ where } i = 1, 2, 3 \ldots n \text{ and } Q(\theta) = 1 - P(\theta)$$

$I_i(\theta)$ is stated as the information at θ of that item. $P_i'(\theta)$ is the first derivative of P with respect to θ. From the equation above, some findings can be concluded: (i) information maximizes when b approached to θ, (ii) information is proportional to item discrimination parameter a, and (iii) c is inversely proportional to information.

Standard error of ability estimation is given as the inverse square of information at that ability level.

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

Good item pools are expected to have items with higher "a", a larger range of "b" values, lower "c" values.

To select an appropriate IRT model, model-data fit must be checked after unidimensionality and local independence are proved to be hold.

Among the model-data fit assumptions non-speeded test administration is another assumption to be met for all dichotomous IRT models, verification of which not so easy. If numbers of students who completed test in different amounts are close to each other, the assumption can be regarded to hold. To use 1P model,

equal discrimination indices assumption should be hold. Investigation of biserial or point-biserial correlations can be helpful for checking that assumption. For 1P and 2P models, minimal guessing should be controlled for they have no chance parameter by, for example, investigating least-able students on hardest items. In addition to them, invariances of item and ability parameters are other assumptions that should be checked.

In a study about IRT, Yıldırım, Comlekoglu and Berberoglu (2003) investigated the fit of items from Private School Entrance Examinations to IRT. Findings of the study stated that data set has meet assumptions of unidimensionality, local independence and low pseudo-guessing and researchers stated that IRT could be used for this examination.

In addition to these studies, there are other studies conducted by Berberoglu (1988), Ertkin (1993) and Caliskan (2000). In these studies, different perspectives of IRT and fit of several data sets to IRT models were investigated.

Berberoglu (1988), in his dissertation, tried to find out potential contributions of Rasch Model to Student Placement Examination conducted in Turkey. Researcher stated if assumption of Rasch Model was hold, advantages of IRT can be helpful for several stages of test development and administration of Student Selection Examination. As a result of the study, Berberoglu reported no differences between scores obtained CCT and IRT Rasch model.

Akyildiz (2003) in his study compared ability estimations obtained from scores of students given Student Selection Examination using IRT and CTT. Researchers compared ability estimations calibrated using 3PL IRT model to CTT ability estimations. Researcher indicated correlations ranging from 0.47 to 0.60 for different subtest of Student Selection Examination and reported to statistically significant differences among correlations.

For assessment of goodness of fit of data to IRT models, several statistics were produced by researchers such as Bock's chi-square (BCHI) (Bock, 1972), Yen's Chi-square (YCHI) procedure (Yen, 1981), and Wright and Mead chi-

square (WCHI) (Wrigt & Mead, 1977). Later, a new index, G^2, was proposed by Orlando and Thissen (2000).

Beside the fit of data to IRT models, fit of examinee responses is a field with increasing interest of researchers. Several statistics were proposed for assessment of person fit. Among them are *l* (Levine & Rubin, 1979), l_z (Drasgow, Levine, & Williams, 1985), *D(θ)* (Trabin & Weiss, 1983) as parametric indices, and *MCI* (Harnisch & Linn, 1981), *U3* (van der Flier, 1980), *ZU3* (van der Flier, 1982) as nonparametric indices. For a very comprehensive study that includes comparison of thirty-six person fit statistics, see Karabatsos (2003).

Also there are commercial and noncommercial computer softwares to assess fit of data to IRT models. Bilog-MG (Zimowski, Muraki, Mislevy & Bock, 1996) is a commercial alternative for assessment of goodness of fit analysis for IRT models. As an alternative to those, IRTFIT_RESAMPLE, a free computer software developed by Stone (2004), can be used for the same purpose. Also Liang, Han and Hambleton (2009) developed a computer software for graphical goodness of fit analysis named ResidPlots-2.

Commercial computer software are also used to generate item and examinee response data with desired statistical properties. One of the alternatives for data generation is WinGen developed by Han (2007). WinGen allows user to define IRT model (dichotomous: one, two, and three parameters; polytomous: partial credit model, generalized partial credit model, Samejima's graded response model, rating scale model, and nominal response model or nonparametric models: kernel-smoothed ICCs, etc.). Also users can select distribution of ability estimation among normal, uniform, b distribution and a normal, uniform, b, or log-normal distribution for item parameters.

2.4 Ability Estimation

Lord (1986) discussed MLE and Bayesian parameter estimation techniques from the perspective of IRT. Lord underlined that Bayesian parameter estimation techniques are better than MLE because Bayesian techniques uses

more information. Another point stated by Lord is that since ability estimation conducted using Bayesian techniques used psychometric properties of sample under investigation, the same response pattern provided by examines may generate different ability estimations unlike MLE which produces always the same ability estimation for the same response pattern. On the other hand, MLE did not produce different estimates based on scale on which ability estimates are put. Stating that Bayesian estimation did not produce divergence, Lord suggested using Bayesian techniques.

Birnbaum (1958) was the first to propose to use maximum likelihood estimates of ability that is MLE ability estimation method. Birnbaum was also the first person who proposed terms item and test information functions. Although Birnbaum dealt with two and three parameter IRT models, using his findings to make generalization for more IRT models, Samejima (1969) proposed a Bayesian estimator. This estimator is based on maximization the posterior density of ability using examinees' responses to items.

But it was the study of Bock and Aitken (1981) to make mathematical calculations possible for using estimator proposed by Samejima who did not further investigated use of that estimator. They proposed the terms MAP and EAP, introducing the calculation techniques for practical test administrations.

Bock and Mislevy (1982) evaluated Bayesian EAP ability estimation method from the perspective of CAT administration. They stated advantages of EAP over MLE and MAP as follows:

- EAP estimates are easy to compute. They do not require long and complex mathematical iterative computations,
- Unlike MAP, they need no derivatives, which makes them free of prior distribution assumptions,
- Unlike MLE, they are always produced. They work well for all-wrong and all-correct situations of examinees,

The advantage that Bayesian methods has over MLE overcomes the problem of nonexistent maximum point. In every circumstance Bayesian methods produce ability estimations therefore they can be used in zero correct, full correct and aberrant response situations (Hambleton, Swaminathan, & Rogers, 1991). But they need complex computational operations to estimate ability than MLE. EAP produces the lowest standard errors (that is, the most reliable estimations) among all ability estimation procedures, but it is biased and needs a priori distribution (generally a normal standard distribution). This bias is the main reason for not being preferred for CAT administration, despite their lower standard errors. Wang, Lau and Hanson (1999) stated that MAP produce less bias, on the other hand it tends to yield higher standard errors.

MAP method is similar to EAP in that it uses prior information about ability distribution of examinees, but it uses mode rather than mean which EAP uses. OWEN method but prior information is updated using a normal distribution. Wang and Wispoel (1988) showed OWEN method had the poorest estimations.

Bock and Mislevy stated (1982) states the differences between MAP and EAP methods as follows: (i) EAP is easier to compute; (ii) EAP is independent of assumptions of distributions defined a priori.

Raîche and Blais (2002), in their paper, proposed using of Bayes EAP method for CAT administration. They stated that using Rasch Model and EAP for ability estimation method, number of items required for a CAT administration was in range between 13 and 40 for achieving a standard error between 0.40-0.20. However, researchers also indicated that this approach generate a bias when there is a significant difference between examinee ability level and a priori ability level. To reduce that bias they suggest using correction methods such as adaptive correction for bias (ACB), adaptive a priori estimate (AAP), and adjusting and adapting the integration interval of the a priori estimate (adaptive integration interval, IN). As a result they reported results in favor of combined use of AAP and IN methods to reduce bias.

Wang (1997) proposed a new expected a posteriori (EAP) estimation method to overcome limitations of existing EAP ability estimation method such as bias in ability estimations. That new index has a flatter prior distribution rather than standard normal used in many EAP estimation sessions. By this way, prior distribution do not indicate examinees' prior ability distribution in opposite of current. Researchers tried different alternatives for prior distribution using simulation techniques and found that beta distribution produces minimal bias, even less than MLE method without losing its small SE advantage over other ability estimation method.

2.5 Test Termination

Lord and Stocking (1987) conducted a study about stopping rule for computer adaptive tests and found that tests with variable length can affect ability estimations negatively, especially if test is short. However, among the other stopping rules, fixed error of SE which can be obtained by variable test length seems the best alternative.

Riley, Conrad, Bezruczko and Dennis (2007) explored effects of using test stopping rules on shortening the number of items given to examinees in CAT version of Global Appraisal of Individual Needs' (GAIN) Substance Problem Scale (SPS) for examines with different ability levels. GAIN includes a number of measurement instruments used in North America. These instruments are designed to be completed 1 to 2 hours and are used to determine level substance abuse treatment. Researchers used a test design to investigate the effects of different test stopping rules on different samples, they defined 0.35 logit for middle range ability levels and for low and high ability levels SEM values they defined are 0.50, 0.60, and 0.75 logits as a result of their study, they stated that relaxing strict test stopping rules for different ability levels make a significant contribution to the reduction rate of items given. They reported 13% to 66% reduction rates.

Babcock and Weiss (2009) conducted a study to investigate performance of different test stopping rules in CAT administration. In their study, researchers

defined several test stopping rules such as SE, fixed number of items minimum information, change in ability level, etc. Also they used item banks with different characteristics to use test stopping rules. Using they simulated 100 examinees for 13 ability points on the continuum of ability. They run fourteen simulations and results indicated that non-fixed number of items in CAT administration performs equally to fixed number of items.

Yi, Wang, and Ban (2001) examined the effect of test termination rules on ability estimation in CAT administration. They focused at three different termination rules: fixed length, SE thresholds and information thresholds. They conducted simulations and reported significant bias in ability estimation related to test termination method selected. They stated that test termination is an influential factor on ability estimation. They suggested that use of SE threshold would decrease the efficiency of the test.

Also a similar situation reported by Simms and Clark (2005) who tried to validate a CAT version of the Schedule for Nonadaptive and Adaptive Personality (SNAP) which is a collection of scales related to personality disorder. When adapting SNAP to a CAT format, they use a two-stage test termination rules: they defined a rule that a minimum number of items given to student and then use either SE threshold or a information threshold (which one achieved first). They observed that 82% of participants were given near to maximum in number, even some participant receive more items than item bank includes. They explained this situation based on poor psychometric properties of some subscales.

On the other hand, Babcock and Weiss (2009), in their research, conducted a comprehensive analysis using item bank with different psychometric characteristics to investigate performances of test termination rules. Given that SE threshold is taken low enough to provide a highly reliable measurement, performance of fixed SE method is no less than fixed length termination rules even though a small number of items are used. They reported that bias stemming from use of SE threshold test termination is a statistical artifact. They discussed the situations potentially cause to that. They suggest that use of a combined test

termination approach combining fixed SE and fixed number of items might be a solution as to select test termination rule.

Wang and Wang (2001) conducted a Monte Carlo study to compare several ability estimation methods. One of the independent variables researchers used is test stopping rule. They adopted two different stopping approaches: fixed test length and fixed test reliability. They used different criteria for each stopping rule and checked their effect to investigate the performance of ability estimation methods. As a by-product, they reported that effect of test stopping rule on ability estimation is more than those of psychometric properties of item bank, especially for MLE.

2.6 Summary

In summary, investigating computer adaptive testing administration from different perspective has became very popular. It is important to note that the results of these studies indicated that CAT is an appropriate technique for measuring individuals with fewer items and higher reliabilities compared to paper and pencil format. On the other hand, there are no studies related to applicability of computer adaptive test applications in Turkey.

CHAPTER 3

METHODOLOGY

In this chapter, methodology of the dissertation is presented. The present study has two phases. First phase includes post-hoc simulation studies based on real examinees' responses. In that phase, using post-hoc simulation techniques based on responses of live examinees' responses to past SSEs, ability estimations obtained from simulations and paper and pencil formats of SSEs were compared. Purpose of the first phase is to find out best CAT testing strategy in terms of ability estimation methods and test termination rules. The principle reason for using real data simulation is that use of generated data (i.e. Monte-Carlo simulations) may not be reflecting examinee's psychometric characteristics and some factors such as speedness and guessing are hard to simulate (Wang, Pan, Harris; 1999). Use of real data (post-hoc simulations) that presents characteristics of examinees is more useful for the purpose of the present study since it uses real responses of live examinees. Simulation application choices an item as that item is given to a student, then computer application checks the response those students gave earlier since responses of that student exist in the database. Then computer picks another item and check response given by student, etc. Simulation phase provides invaluable information about applicability of CAT administration of SSE science subtest based on real responses such as correlations between ability recovered from P&P and CAT, number of items given in CAT compared to P&P, distribution of standard errors for CAT, etc.

At the second phase, a CAT application to real examinees based on an item bank including previous SSE science subtest items which has almost 242 items was conducted. Then ability estimations of real students from P&P test and

CAT administrations were compared. Ability estimation method and test termination rule used in live CAT administration was obtained from post-hoc simulation phase. A live CAT administration was conducted. In post-hoc simulation phase there is no effect on examinees arise from CAT administration since response patterns used for simulations were provided for P&P test administration. Examinees may develop different attitudes in front of computers when they are given CAT format of SSE science subtest.

Software used both in post-hoc simulation and live testing application phase were developed by researcher using Delphi platform using Object Pascal.

3.1 Sample of the Study

Data sets used in the present study were obtained from Student Selection and Placement Center. Electronic files include all students' responses in dichotomous format who take SSE for years 2003, 2005, 2006, and 2007. Data sets for the years 2000, 2001, 2002, 2004 only includes item parameters estimated by Classical Test Theory.

Data sets for years 2005 and 2006 were used from calibration and post-hoc simulations phases. For formation of item bank of live CAT administration, all items were used.

For simulations, different high school types represented different examinee groups that are potential CAT test-takers and also different cognitive ability groups. Anatolian high schools have students with higher ability levels since they use a selection procedure. State high schools accept any student without using any selection criteria and private high schools are paid-schools. These three school types follow the same science curriculum therefore results obtained from CAT and P&P SSE science can be compared across schools types. Median values for the correct response to SSE 2007 Science subtests are 7, 35, and 26 for state, Anatolian and private high schools, respectively. State high school examinee group represents 37.19% of whole SSE test takers. Anatolian and private high schools include 9.51% and %2.0, respectively. Reason for

selecting state high schools is that they represent the largest examinee groups in SSE and Anatolian and private high schools has smaller examinee groups, however including them to the present study is expected to yield findings as to applicability of CAT to different examinee groups no matter their ability levels.

For post-hoc simulations data sets for two different years were used to investigate the effect of number of science items in respective SSE science subtest (2005 science subtest includes 45 items, and 2006 set has 30 science items). Descriptive statistics of total science scores for the school types in the study are given in Table 3.1.

Table 3.1 Descriptive Indices for Science Total Scores for Different Samples

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
Mean	9.61	31.72	24.02	5.67	15.60	11.32
Median	7.00	35.00	26.00	5.00	17.00	11.00
Mode	0	40	1	1	1	1
S.D.	9.53	10.50	13.12	5.13	8.33	8.68
Skewness	1.10	-1.45	-0.36	1.04	-0.50	0.29
SE of Skew.	0.01	0.02	0.03	0.01	0.01	0.03
Kurtosis	0.51	1.64	-1.07	0.75	-0.92	-1.16
SE. Kurtosis	0.01	0.03	0.07	0.01	0.03	0.06
Range	45	45	45	30	30	29

As can be seen from the table, median values for total score are different across different school types, indicating different ability levels for examinees on

different school types. State schools have students from lowest part of the ability continuum, on the other hand Anatolian high schools using student selection procedures have much more higher correct response, which means higher ability levels for students.

Calibrations were separately conducted three different school types using randomly selected 1000 students for each sample. Bilog-MG was used for calibrations (Zimowski, Muraki, Mislevy & Bock, 1996).

Sample of the live CAT phase included 33 examinees that were taken SSE 2007 were randomly selected from English Language Preparatory school student at Middle East Technical University. The reason that preparatory class students selected is that after SSE, they were received no additional education that can affect validity of CAT application. All participants received 30 items from P&P SSE 2007 science subtest. Median of correct responses is 28 out of 30 corresponding to 93.33% of items. Medians of ability and SE estimations were found to be 1.21 and 0.22, respectively.

20 (60.6%) of the examinees who participated to live CAT administration are male and 13 (39.4%) are female. Number of participants who have taken a computer-related course is 19 (57.6%). 14 (42.4%) of the participant have not taken any course related to computers. As can be expected from the higher ability levels of the participants (Table 3.1), they are graduated from Anatolian high schools which accept students via a selection procedure. 25 (75.8%) students are from Anatolian high schools and the rest (8, 24.2%) are graduates of other school types.

Table 3.2 shows that examinees who were given live CAT are from high ability groups of P&P SSE participants

Table 3.2 Live CAT Examinee Characteristics

	P&P		
	Ability	SE	# of Items
Mean	1.77	0.47	
Median	2.22	0.50	
sd	1.22	0.22	30
Minimum	-0.99	0.17	
Maximum	3.55	0.99	

Ability distribution of real CAT application can be seen in Figure 3.1.

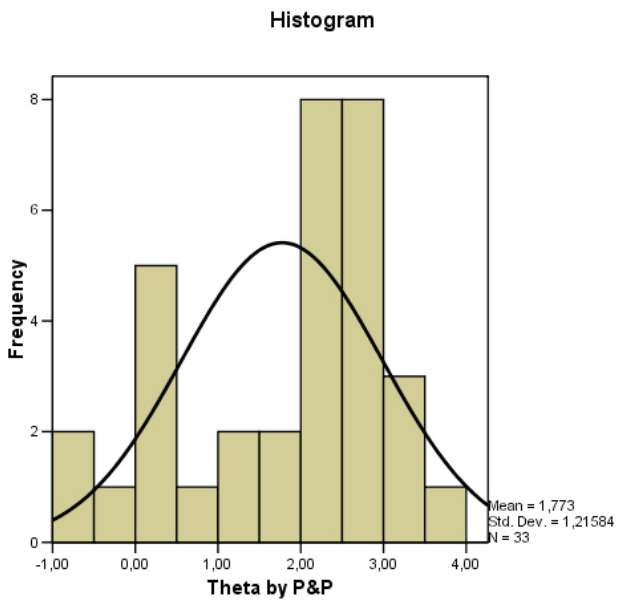

Figure 3.1 Ability Distribution of Participants of Live CAT

As can be seen from descriptives of ability estimations participant of CAT session are high ability examinees.

3.2 Assessment of Model-Data Fit

Before proceeding to simulation and real examinees phases, assessment of model-data fit must be investigated to check the data sets are appropriate for benefiting from the advantages of the IRT.

Assessment of model-data fit includes three stages (Hambleton & Swaminathan, 1984): (i) checking model assumptions, (ii) checking expected model features, and (iii) checking model predictions. First stage includes checking unidimensionality, equal discrimination indices, minimal guessing, and non-speeded test administration. In the second stage, invariance properties verified. Invariance of ability estimations and invariance of item parameters are two features that should be obtained when using IRT models. And for third stage, model predictions are investigated to assess the deviations of estimated values from actual values.

For the assessment of model-data fit, students with all-blank zero responses patterns (those with all responses are blank) for science subtest were excluded from the analyses. Resulting data sets includes examinee provided at least one response. Another criterion for selecting student to data sets is that students' educational position in the system. Only student who are at the last grade of the high school were included to eliminate the effect of out-of-school instruction after graduation such as private tutoring, etc.

Assessment of model-data fit data sets for the years of 2005 and 2006 SSE science subtest. SSEs for different years are given by students who have similar cognitive characteristics across years and items in the test do not significantly differ from year to year. Therefore selected IRT models were accepted to be hold for other years.

3.2.1 Model Assumptions

Unidimensionality assumption of IRT means that examinee's performance can be depicted by one single dimension. And it is a strict assumption for all dicthotomous IRT models to be met. This, however, is rarely if ever met in practice (Hambleton & Swaminathan, 1985). This assumption can be approximated by assessing the ratio of first to second eigenvalues, which is an index of the strength of the first dimension of the data (Reise & Waller, 1990). This implies that first factor explains a large proportion of the total variance, which means that assumption of a dominant factor has been met. TESTFACT (Wilson, Wood & Gibbons, 1991) were used to conduct factor analyses since it uses tetrachoric correlation. First five eigenvalues for each school types are given in Table 3.3.

Table 3.3 Eigenvalues for Each School Types for Different Years

	2005			2006		
#	State	Anatolian	Private	State	Anatolian	Private
1	21.17	24.37	26.70	10.65	15.94	17.78
2	1.78	1.38	1.56	1.55	1.04	1.02
3	1.45	1.13	1.13	1.23	0.97	0.97
4	1.15	1.01	0.87	0.93	0.89	0.84
5	0.83	0.84	0.80	0.89	0.82	0.74
1 / 2	11.89	17.61	17.10	6.85	15.33	17.40

Ratios of first eigenvalues to the second ones indicated that tests are unidimensional with a strongly dominant first factor.

Local independence means that after conditioning on ability, examinees' responses to the items on the test are likely to be independent (Hambleton et al, 1991). In general, when the unidimensionality is met, assumption of local independence is said to be met. On the other hand, even assumption of unidimensionality is met, local independence can not be satisfied (Lord, 1980). Investigation of inter-item correlations among subgroups in terms of ability levels can be used for checking local independence. Table 3.4 shows the means of inter-item correlations for whole groups, and restricted ability subgroups (low and high ability groups).

Table 3.4 Inter-Item Correlations for Subgroups

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
whole	0.254	0.299	0.364	0.167	0.314	0.358
low ability	0.081	0.097	0.08	0.068	0.089	0.089
high ability	0.015	0.037	0.039	0.011	0.015	0.022

Sharp drops on the correlations for low and high ability groups indicate that local independence was satisfied for each sample. That means that SSE science subtest measure one single common trait of examinees.

Assumption of non-speeded test administration, another assumption essential to all IRT models, can be checked by investigating percentages of missing responses for last items were examined. However, there are 10 different booklets which include the same items but in different orders and with choices with reordered. Therefore last items for one examinee are different for each

examinee. The data sets obtained Student Selection and Placement Center does not have any variable to define booklet information. So it is not possible to check this assumption because last items are different for examinees.

Assumption of equal discrimination indices, required for 1PL, was checked by investigating the classical item discrimination indices obtained by ITEMAN (2010). As seen on the Table 3.5, item discrimination indices are not homogenous, so it can be concluded that assumption of equal discrimination indices was not met. Since this assumption is required for 1PL Model, the model-fit for 1PL model was considered not to be satisfied.

Table 3.5 Descriptives for Item Discrimination Indices for SSE Science Subtest

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
Median	0.697	0.703	0.772	0.531	0.7125	0.81
Range	0.477	0.654	0.396	0.56	0.731	0.612
Minimum	0.444	0.430	0.554	0.222	0.185	0.317
Maximum	0.921	1.084	0.950	0.782	0.916	0.929

Minimal guessing assumption, for 1PL and 2PL, can be verified by demonstrating low ability students showed low performance on hard items. Table 3.6 shows the classical item difficult parameters (p) and percentage of correct responses (%) for each sample. Since SSE is a five-option test, proportion of correct responses on the most difficult items should be lower than 0.2 to satisfy the assumption of the minimal guessing. p in the Table 3.6 indicated item difficulty index of classical test theory.

Table 3.6 The Correct Responses on the Most Difficult Items

2005						2006					
State		Anatolian		Private		State		Anatolian		Private	
p	%	p	%	p	%	p	%	p	%	p	%
5.5	3.4	34.9	1.2	20.5	1.5	5.6	4.0	6	1.2	3.9	1.7
9.7	4.2	36.4	2.9	28.8	5.8	5.7	4.5	26	1.4	18.5	2.3
10.3	7.6	39.5	5.3	29.1	7.3	9.5	6.4	32.9	2.4	21.3	3.0
10.6	5.7	43.1	4.2	29.5	5.5	10.3	5.9	35.8	2.2	21.4	2.2
10.6	6.9	44.6	4.5	30.2	4.4	10.8	7.6	37.7	12.9	27	6.2

All the percentage values for each item on the samples are lower than 0.2, indicating the minimal guessing is met.

Findings of checking model assumptions revealed that data sets are appropriate for 2PL and 3PL, not for 1PL since equality of discrimination indices assumption did not hold.

3.2.2 Expected Model Features

All IRT models should have two expected model features. First one is invariance of ability estimates which means that ability estimates of examinees are independent of any particular sets of items calibrated for population. Second feature is invariance of item parameters which states that item parameters are independent of groups of examinees from population. (Hambleton & Swaminathan, 1985)

To check invariance of ability estimates, data sets were divided into smaller samples in item numbers such as even/odd-numbered items, hard/easy items and ability estimates for each examinee are calculated for each item groups.

Table 3.7 shows the correlations between ability estimates for even- and odd-numbered items for each sample.

Table 3.7 Correlations between Ability Estimates for Even/Odd Numbered Items of SSE Science Subtest

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
Even/Odd	0.868	0.754	0.996	0.877	0.872	0.913

High correlations between ability estimates indicate that assumption of invariance of ability estimates are satisfied.

Invariance of item parameters is another expected model feature that should be checked. For each of different samples correlations estimated from low and high ability groups were compared and results were presented in Table 3.8. Low and high ability groups were obtained from examinees constituting lowest and highest 10% of whole examinee group.

Table 3.8 Correlation between Item Parameter Estimates for Low/High Ability Groups of SSE Science Subtest

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
a	-0.380	0.504	0.350	0.33	0.235	0.401
b	0.451	0.864	0.882	0.359	0.811	0.906
c	0.267	*	-0.093	-0.051	*	-0.480

* means that no correlation coefficient were estimated

Some correlations were found to be lower than expected. Reason for lower correlations than expected was given by Stocking (1990) who stated that heterogeneous samples are needed to obtain stable estimations of item parameters. Since school types included in the present study were calibrated separately, samples could be regarded homogenous and unstable item parameters can be explained in that way.

3.2.3 Model Predictions

As a third step of model-fit analysis, the model predictions were checked.

To assess model predictions several ways can be followed such as likelihood-based fit indices (Yen, 1981; Bock, 1972; McKinley & Mills, 1985), residual analysis, graphical analysis (Ludlow, 1986; Hambleton & Swaminathan, 1985; Wainer & Mislevy, 1990). Also there are softwares developed by several researchers to assess item-fit (IRTFIT_RESAMPLE by Stone (2004), EO-FIT by Ferrando & Lorenzo-Seva, (2004)).

Likelihood-based fit indices can generally be obtained by (i) estimate and item parameters from a dataset, (ii) sort examinees by their estimates, (iii) form subgroups of the sorted examinees, (iv) calculate the proportion of examinees in each subgroup who answered correctly/incorrectly for each item, and (v) compare these "observed" proportions with those predicted by the model using a -2-like statistic and/or a graphical representation (Ankenmann, 1994).

Likelihood indices have some well-knows limitations such as dependency to sample size and number of intervals used in estimation of parameters.

Though Bilog-MG produces χ^2 estimates for assessment of model predictions, the fact that it is highly sensitive to sample size, these values can be misleading. Instead of likelihood indices, visual interpretations of ICCs produced by BILOG-MG were preferred.

As a result of investigation of ICCs for 2PL and 3PL – equal discrimination indices assumption does not hold, therefore 1PL was eliminated –

3PL was chosen for the analyses in the rest of the present study. Some of items who do not produce a good fit were excluded from the rest of analyses.

Based on the findings of the assessment of model-data fit analyses, (i) 3PL model was selected for IRT analyses, and (ii) some of non-fitting items were excluded. In post-hoc simulation and live test phases, 3PL logistic (D=1.7) model is used. ($P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}}$ where i is i[th] item)

Ability estimation of examinees used in calibration phase was defined to be the same with ability estimation method of post-hoc simulations. Ability distribution after scoring the examinees was transformed to have mean and standard deviation 0 and 1, respectively.

3.3 Equating of Test Scores

Putting items on a common scale is an important issue in developing item pools. To this end, (i) common items nonequivalent groups design or (ii) random groups design can be used (Kolen & Brennan, 1995). None of the these designs provide a complete solution to the issue of putting parameter estimates to a common scale since neither there is no common items in the forms across years nor any evidence as to equivalence of the groups taken tests in different years.

SSEs for different years are taken by students who have similar cognitive characteristics across years. Also item format and measured traits in the tests do not significantly differ from year to year.

Based on that, item parameters and ability estimations were accepted to be a common scale across years.

3.4 Ability Estimation

There are two ability estimation methods used in the present study given by the following equations.

MLE:

$$\frac{\partial \ln L(u \mid \theta)}{\partial \theta} = \sum_i \frac{P_i'(u_i \mid \theta)}{P_i(u_i \mid \theta)} = \sum_i \frac{(u_i - P_i)P_i'}{P_i Qi} = 0$$

EAP:

$$\widehat{\theta} \equiv E(\theta \mid u) = \frac{\sum_{k=1}^{q} X_k L(X_k) W(X_k)}{\sum_{k=1}^{q} X_k L(X_k)}$$

3.5 Post-Hoc Simulations

3.5.1 Post-Hoc Simulation Working Principle

Post-hoc simulation uses real examinee responses and conducts a CAT simulation for each examinee using the item responses as if examinee gives the responses that provided in paper and pencil test in a CAT session. By this way, reduction rate in the items of CAT compared to P&P can be determined.

Working principle of post-hoc analyses is as follows:

- Students' responses are obtained in dichotomous format in a paper and pencil test.
- Items are calibrated to obtain IRT parameters of the item in that test.
- When simulation starts for first examinee, computers pick an item and checks the responses of the examinee to that item as if the item is asked to that examinee, since examinee took that exam before in P&P format.
- Based on that response, computer picks another item based on predefined item selection rules and checks examinee's response to that item as if examinee provides a response in front of a computer.

- Computer goes on to pick items until predefined test stopping rules is hold. This is repeated a number of examinees.
- At the end of the simulation, reduction rate of items by CAT format can be investigated by checking correlations among ability estimates obtained from CAT and P&P format.

3.5.2 Software for Simulation

To conduct post-hoc simulation researcher developed a computer application named CATSIM. This is a computer application similar to POSTSIM, a commercial post-hoc simulation package, and provides users with an opportunity to conduct a post-hoc simulation. Program, screenshot of which can be seen in Figure 3.2, allows users to provide (i) starting line of post-hoc analyses in case some may want to skip some line and start at another line, (ii) number of examinee that will included to the simulation and (ii) stopping rule (fixed number of items or a threshold for standard error), and (iv) ability estimation method (MLE or EAP). CATSIM uses Maximum Information method for item selection. At the end of the simulation, CATSIM creates three output file. One file includes detailed CAT progress for each examinee (examinee ID, numbers of items given, ID of the items given, examinee responses, ability estimation and standard error after each response). Second output file provides a summary of the first file (examinee ID, numbers of items given, ability estimation and standard error for each examinee). And last output file includes number of items each item is used, that is, item exposure rates for items. Program needs two input files: (i) item responses including IDs for each examinee, number of characters for ID can be determined by user, and (ii) an item bank including item parameters for running. There is no limitation for numbers of examinee and items rather than computer limitations. Sample outputs can be seen in Table 3.9 and Figure 3.3.

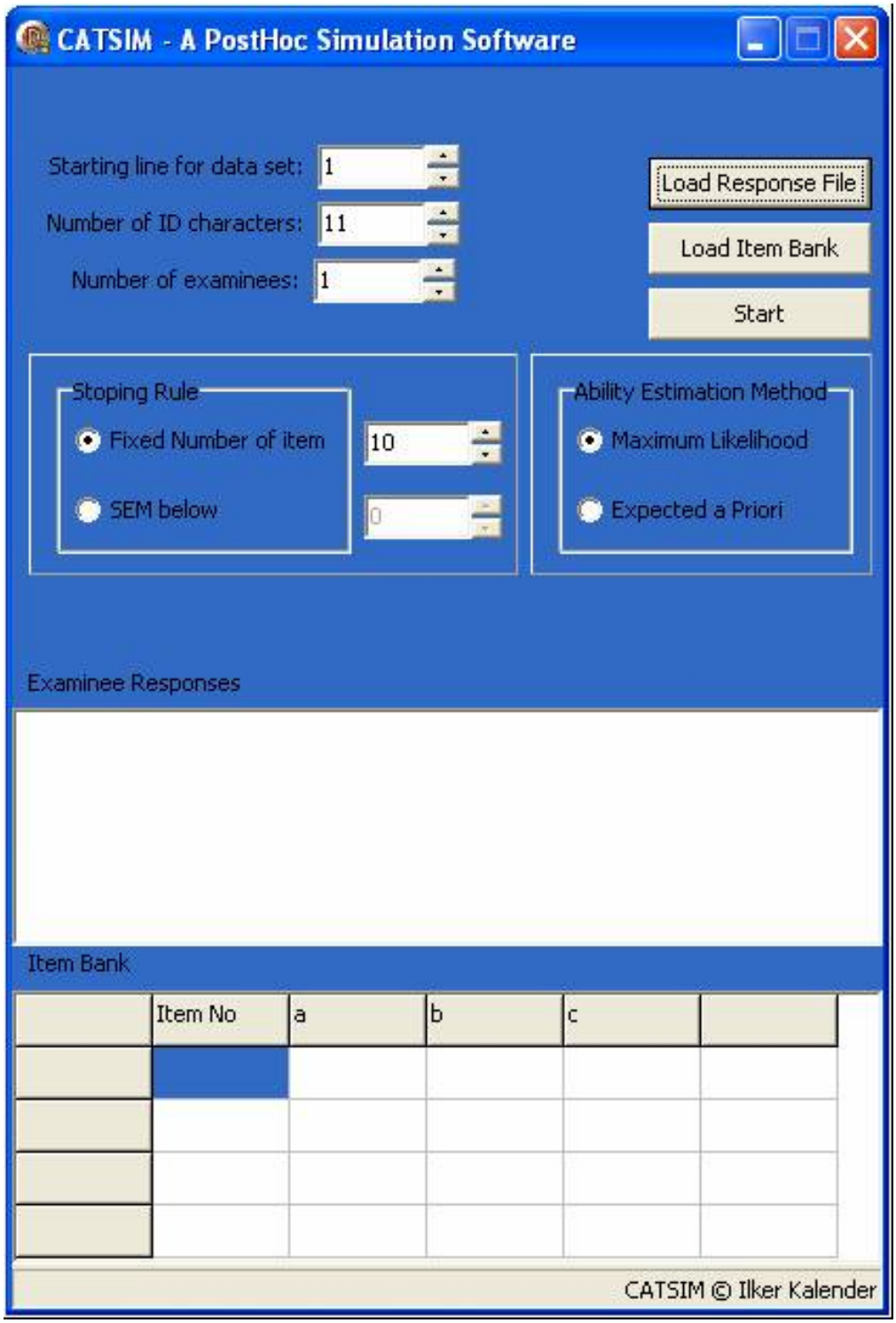

Figure 3.2 CATSIM User Interface

Table 3.9 Sample Output of CATSIM

---------- CATSIM ----------

Post-hoc Simulation Application for CAT

Date: 99.99.2099 **Time:** 13:44:07

Starting line data set: 1

Number of ID characters: 11

Number of examinees: 30

Method of ability estimation: Maximum Likelihood

Method of item selection: Maximum Information

Stopping rule: Fixed Standard error of estimation (0.30)

ID	Order	Item #	#of T	#of F	Theta	SE
50001	1	1	0	1	NA*	NA*
50001	2	19	0	2	NA*	NA*
50001	3	18	1	2	-0.9	0.9363
50001	4	34	1	3	-1.43	0.638
50001	5	17	1	4	-1.7	0.575
50001	6	28	2	4	-1.46	0.4312
50001	7	13	3	4	-1.3	0.3733
50001	8	14	3	5	-1.42	0.3519
50001	9	31	3	6	-1.57	0.3396
50001	10	22	3	7	-1.67	0.3292
50001	11	24	3	8	-1.77	0.3233
50001	12	8	4	8	-1.71	0.3043
50001	13	44	4	9	-1.76	0.2986

*Examinee give false responses to first and second items, and since MLE was used for ability estimation method, no ability estimation produces until one correct/one false response pattern was obtained.

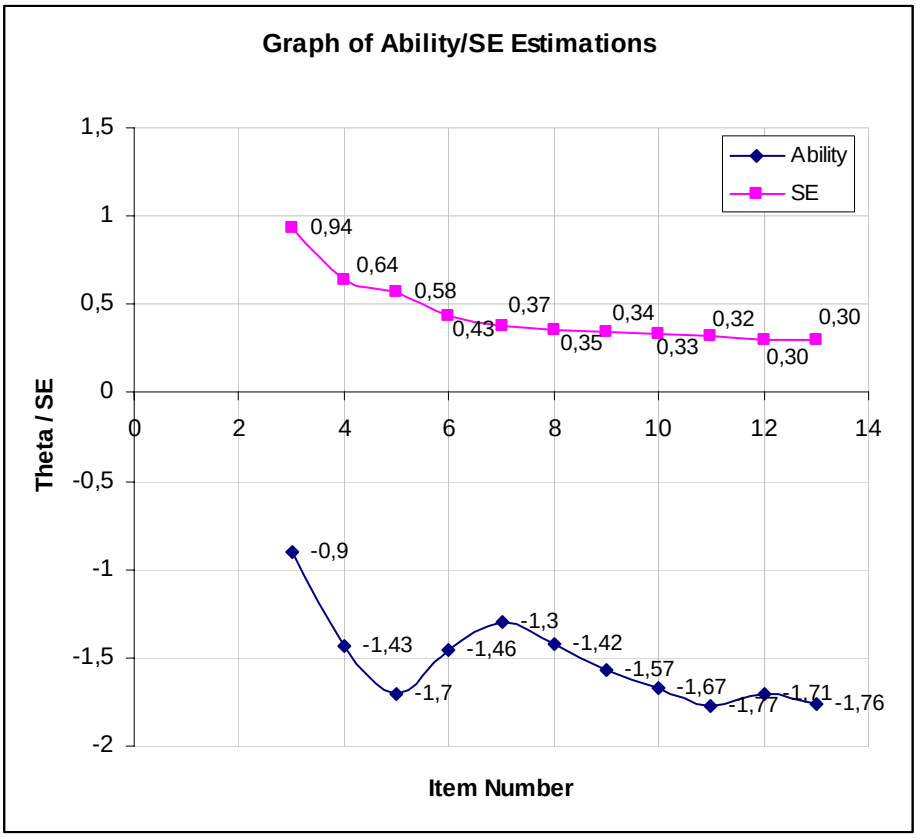

Figure 3.3 CATSIM Sample Output Plot

3.5.3 Item Pool Characteristics

Based on the assessment of model-fit, 3-parameter model were decided to use. Item calibration BILOG-MG was used. Table 3.10 shows mean and standard errors of item discrimination (a), item difficulty (b) and pseudo-guessing parameter (c).

In the calibration phase, maximum number of EM cycles and maximum number of Newton cycles were set to 40 and 10, respectively. As convergence

criterion 0.01 was used. Chi-square fit indices for the item parameter estimates were indicated some of the items did not fit to the model used. However, on P&P examinations examinees were given all items in the test independent of whether some items are faulty. To keep parallel between the real and simulation tests, no items were excluded based on model-fit analyses (Non-fitting items to 3PL were excluded from live testing phase item bank.).

Table 3.10 Means of IRT Item Parameter Estimates with SEs for Post-Hoc Simulations

	a	b	c
State 2005	2.17	1.21	0.01
State 2006	1.63	1.65	0.01
Anatolian 2005	1.65	-0.78	0.01
Anatolian 2006	2.13	0.11	0.01
Private 2005	1.38	-0.02	0.01
Private 2006	1.57	0.57	0.02

Item parameter estimates indicate different school types in the present study have different item difficulty means and this proves that different school types represent different ability groups.

3.5.4 Simulation Design

Working principle of the post-hoc simulation was explained above. Here design of the simulation will be explained.

For simulation phase, different samples were formed using SSE data sets of science subtest. To represent different ability groups, three different school

types (state, Anatolian and private) were included to the study. Also to investigate the performance of recovery of ability estimations using CAT, tests belonging two different years for each school type were also selected. In year 2005 science subtest is 45 items, and in year 2006 it was 30 items. Test content is the same for both years. Therefore six samples were obtained; 3 school type (state, Anatolian, private) x 2 different test length (45 items for SSE 2006 and 30 items for SSE 2005). By this way, investigation of performance of CAT with different cognitive levels and test length using real examinee data became possible. For each of these samples 5000 student were randomly selected for different CAT testing strategies.

Bilog-MG calibration was run for each school types separately for two years of 2005 and 2006 which have test lengths of 45 and 30, respectively. Convergence criterion was set to 0.01. Maximum numbers of EM and Newton cycles were defined as 40 and 10, respectively. 15 quadrature points were used for each calibration. Calibration phase was conducted to select proper IRT model for the present study.

Examinee's full test ability scores were estimated using BILOG-MG by both MLE and EAP. Sample distributions of both examinees' abilities estimated P&P and CAT on each sample were set to have a mean of zero and a standard deviation of 1. Table 3.11 shows the ability means estimated by 3PL model for school types for both years.

Table 3.11 Descriptive Indices for Ability Estimations

	2005			2006		
	State	Anatolian	Private	State	Anatolian	Private
Mean	0.082	1.753	1.201	0.067	1.321	0.783
sd	0.902	0.869	1.030	0.922	1.183	1.289
Variance	0.813	0.756	1.061	0.850	1.398	1.661
Reliability	0.950	0.947	0.968	0.892	0.953	0.955

Testing strategies for the simulations are (i) ability estimation methods – MLE and EAP and (ii) test stopping rules – fixed SE and fixed test lenght.

For MLE, to obtain one correct/one false pattern as soon as possible the following design was applied: First item is selected among 5 items with moderate difficulty. And based on the response of the examinee, next item is selected among the hardest or easiest 7 items. Therefore 70 (2 x (5 x 7)) different test starting pattern exist. By this way, examinees were forced to provide one correct/one false response pattern required for MLE. Since there is no requirement for EAP to make ability estimation, no predefined forcing mechanism was used.

If an examinee does not provide that required pattern for MLE for 8 items, this examine is marked with diverging test termination because its ability estimates diverges to infinity and simulation stops for examinee.

On the other hand, in terms of test stopping rules, different parameters were applied. For threshold of SE, five different levels were stated. These levels are 0,50; 0,40; 0,30; 0,20 and 0.10 which correspond to CTT reliabilities 0,75; 0,84; 0,91; 0,96 and 0,99, respectively.

Transformations were conducted using the formula proposed by Lord and Novick (1968).

$$SE = \sqrt{1 - r}$$

For fixed SE test termination rule an additional rule was also defined. If an examinee did not achieve to a SE of 0.30 after 45 (SSE science subtest 2005) and 30 (SSE science subtest 2006) items, test was terminated. This was done because if more items than P&P SSE science subtest for an examinee's ability estimation, CAT administration is of little use.

For all post-hoc simulations, item selection was based on Maximum Information. This rule is based on selecting the items with the highest information at that ability level..

Since 3PL logistic model was used for the present study, IIF is given as follows

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

Fixed item stopping rule were studied with three different levels; 23%, 33% and 55% of the full test length. For year 2005, 23%, 33% and 55% of the full test correspond to 10, 15 and 25 items, respectively. For year, 2006, 23%, 33% and 55% of the full test correspond to 8, 10 and 17 items, respectively.

All examinees who replied at least one response to SSE Science subtest were included in random sampling for the simulations. And starting points for ability estimations on simulations were set to 0 for all examinees. Omitted items were recoded as wrong responses.

For EAP estimation method, a correction formula was applied. To obtain a corrected estimate of theta, original theta was divided by 1 minus square root of SE estimate of the examinees.

$$\theta_{corrected} = \frac{\theta}{1-\sqrt{SE}}$$

3.6 Live Testing CAT

After conducting simulation, a live CAT testing was administered to a group of examinees to observe relationship between ability estimates from CAT and P&P SSE science subtest.

By live CAT testing, examinees were given CAT format of SSE science subtest and their responses were for a CAT administration unlike post-hoc simulations in which examinees responses to P&P SSE science subtest were used.

3.6.1 Software for Live CAT

Researcher also developed a computerized adaptive testing application using Delphi platform with object Pascal. The application implements a CAT with a pre-calibrated item bank and options for ability estimation (Maximum Likelihood or Expected A Posteriori) and stopping rule (fixed item number or a threshold for standard error) that can be selected by user. Program reports the details of the testing process for each examinee and records them in a file.

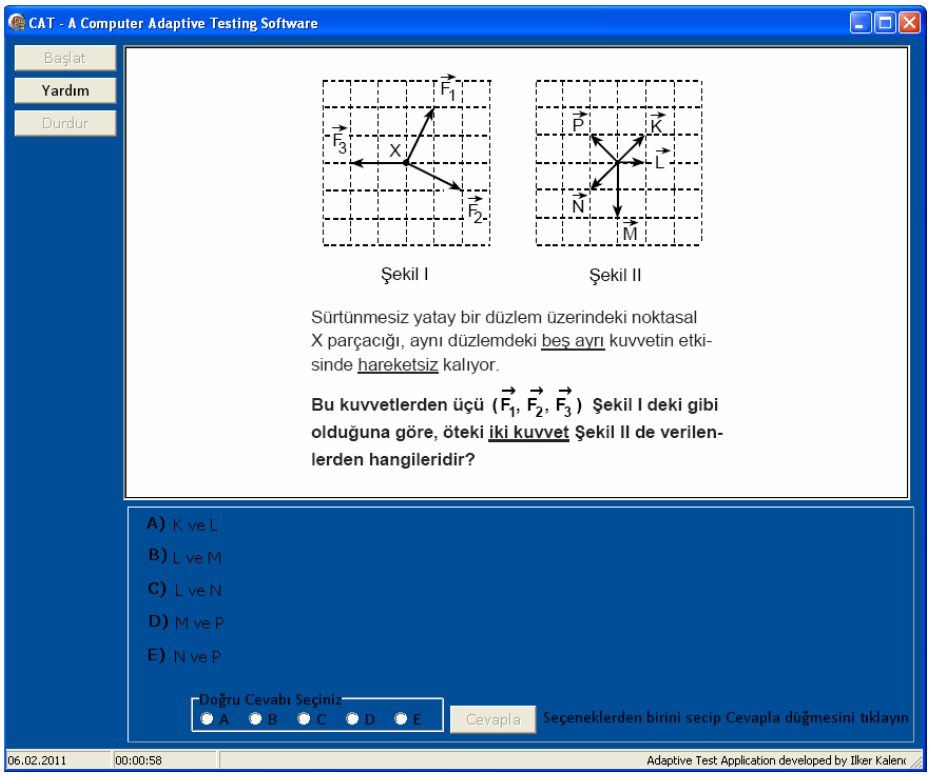

Figure 3.4 Live CAT Software Interface

Sample output for real examinee CAT application can be on the Table 3.12.

Table 3.12 Live CAT Sample Output Plot

Surname, Name	***01.01.2099***										
	Total Test Time: 00:09:11										
Test Progress Report											
Time	**Order**	**Item#**	**a**	**b**	**c**	**Key**	**Choice**	**TF**	**#ofT**	**#ofF**	**Theta**	**SE**
22:51:33	1	4	1.376	0.161	0.200	1	1	1	1	0	NA	NA
22:51:36	2	9	1.351	0.36	0.200	1	3	0	1	1	0.10	0.7922
22:51:37	3	48	1.267	0.485	0.004	4	2	0	1	2	-0.20	0.7600
22:51:38	4	26	1.296	-0.395	0.000	3	4	0	1	3	-3.00	7.9470
22:51:39	5	58	1.085	-2.581	0.200	5	2	0	1	4	-3.00	1.5317
22:51:40	6	59	1.088	-2.106	0.200	5	3	0	1	5	-3.00	1.2839
22:51:41	7	8	1.229	-1.825	0.200	1	1	1	2	5	-3.00	1.2100
22:51:42	8	28	1.349	-1.691	0.000	3	5	0	2	6	-3.00	1.0424
22:51:43	9	24	1.254	-1.615	0.000	2	4	0	2	7	-3.00	0.9388
22:51:44	10	31	1.351	-1.610	0.000	3	3	1	3	7	-2.32	0.5133

3.6.2 Item Pool Characteristics

For real examinee CAT application, data sets from years 2001, 2002, 2003, 2004, 2005, 2006 and 2007 were obtained from Student Selection and Placement Center. 2003, 2005, 2006 and 2007 data sets were in raw data format which include responses of each examinee to each item in the science subtest. On the other hand, 2001, 2002, 2004 data sets were not provided as raw data, rather these data sets included classical item parameter estimates such as item difficulty index, item discrimination index, etc.

Lord and Novick (1968) provided transformation formulas to obtain IRT parameters from CTT item discrimination and item difficulty parameters. Also Gelbal (1994), in his dissertation, compared item parameter estimates of CTT and IRT using transformation formulas proposed by Lord and Novick (1968).

$$\alpha_i \cong \frac{r_{bis}}{\sqrt{1 - r^2{}_{bis}}}$$

where r_{bis} is biserial correlation

and

$$\beta_i \cong \frac{z_i}{r_{bis}}$$

where $z_{is} \cong \alpha_i(\theta_s - \beta_i)$ with i^{th} item and examinee s

This formula set does not include transformation for pseudo-guessing parameter (c). Therefore missing c parameter for each item is set to the mean of c parameters obtained items with raw item information (0.018, mean of c values for 2003, 2005, 2006 and 2007).

Some of items were excluded from the item pool due to low classical item discrimination, IRT fit indices indicating no-fit to the model, and unexpected IRT parameters after using transformation formulas. At the end, a total of 242 items were remained in the item pool.

Item parameters obtained using 3 PL model are given in Table 3.13.

Table 3.13 IRT Item Parameter Estimates for Live Testing CAT

	a	b	c
Mean	0.99	0.89	0.02
Median	0.88	0.82	0.02
Mode	0.63	0.00	0.02
Std. Deviation	0.50	0.73	0.02
Variance	0.25	0.53	0.00
Range	3.44	4.66	0.16
Minimum	0.26	-1.60	0.00
Maximum	3.70	3.06	0.16

Figures 3.5 to 3.7 show the histograms of item parameters a, b and c estimated using 3 PL model.

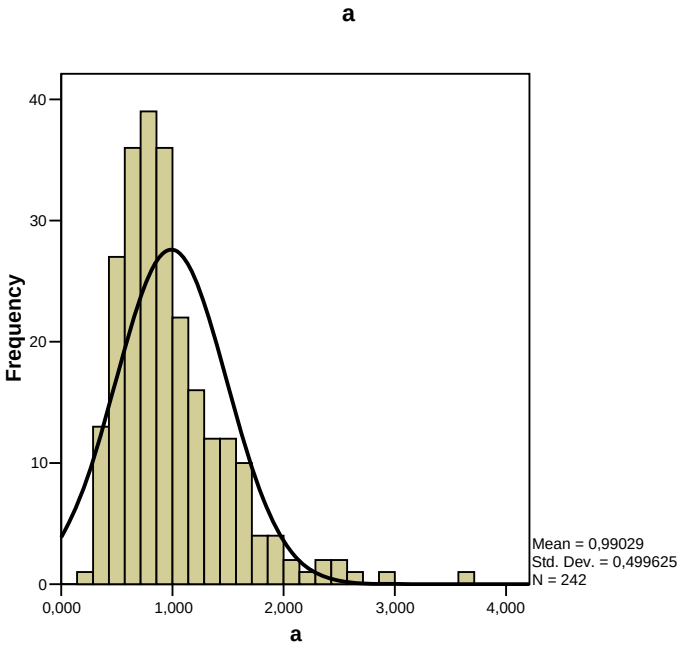

Figure 3.5 Distribution of a Parameter

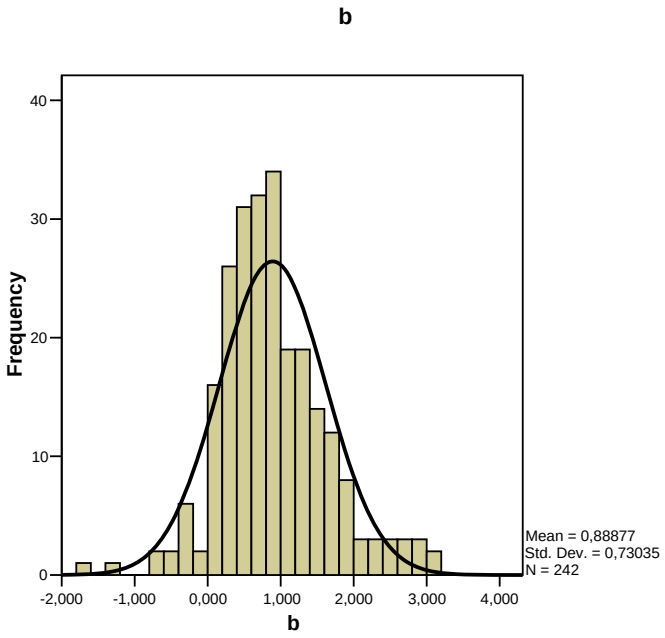

Figure 3.6 Distribution of b Parameter

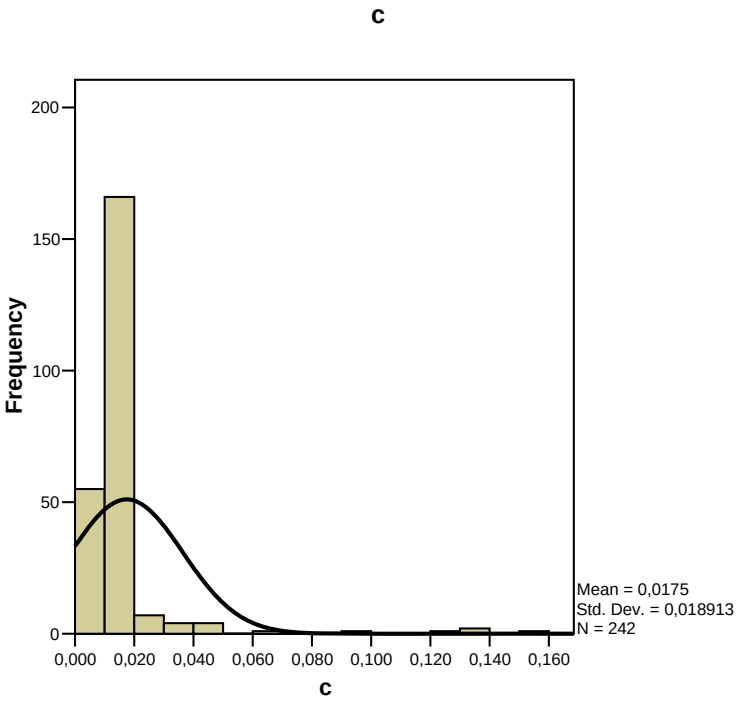

Figure 3.7 Distribution of c Parameter

3.6.3 Procedure

Participation to the live CAT study was completely voluntarily. Examinees were all taken the CAT at the same place and the same conditions. Also the participants were kindly asked to provide their Turkish Republic ID number for obtaining their SSE scores from Student Selection and Placement Center data sets.

Thirty three examinees were participated to CAT administration that were set up with EAP estimation method and fixed SE test termination rule. Threshold for SE was taken to be 0.30 to provide a highly reliable testing session. Using EAP estimation method provided ability estimations for all examinees even if they provided perfect zero/full response patterns.

Like in post-hoc simulation phase, item selection method is based Maximum Information. Moreover, skip or moving along the items is not allowed.

The additional test stopping rule defined as maximum number of items in P&P SSEs was also applied for live testing CAT. Since in SSE 2007 there are 30 items, for any examinee given 30 items without obtaining a SE estimate below 0.30, the CAT administration was terminated by computer.

CHAPTER 4

RESULTS

In this chapter, results of the study are presented. As stated before, the present study has two phases: (i) post-hoc simulation in which ability levels of examinees were estimated using real examinees' responses to items in P&P format of SSE science subtest and (ii) live CAT administration including real examinees. Simulations include using the examinees' responses to simulate CAT sessions as if examinees were given a CAT using an item bank that were constituted items of P&P SSE. In this phase, two ability estimation (MLE and Bayesian EAP) method and two test termination criteria (fixed test length and fixed SE) were defined for examinees from three different school types (state, Anatolian and private high schools). At the end of the post-hoc simulations, the best CAT administration strategy was determined and using that strategy a real CAT administration was conducted using real examinees.

3PL IRT model was used for calibration of items. Ability parameter estimation method in calibration phase was selected to be parallel to estimation method used in simulations.

4.1 Simulation Studies

Results of the simulations are given in this section. As stated before, simulations were designed based on different ability levels (school types), and different test lengths considering different ability estimation methods and test termination rules. Table 4.1 shows the correlations between abilities estimated by P&P and CAT formats of SSE science subtest. All correlations are significant at the 0.05 level of significance.

Table 4.1 Correlations of Ability Estimates between P&P and CAT

Ability Estimation Method	Test Length	School Type	Threshold of SE					Fixed Item*		
			< 0.50	< 0.40	< 0.30	< 0.20	< 0.10	23%	33%	55%
MLE	45	State	0.594	0.636	0.704	0.743	0.578	0.775	0.755	0.789
		Anatolian	0.677	0.802	0.890	0.976	0.781	0.916	0.956	0.983
		Private	0.630	0.733	0.844	0.854	0.543	0.834	0.889	0.942
	30	State	0.529	0.594	0.712	0.803	0.793	0.659	0.736	0.822
		Anatolia	0.665	0.751	0.840	0.934	0.979	0.928	0.968	0.989
		Private	0.761	0.854	0.916	0.961	0.976	0.887	0.937	0.974

* For SSE 2005, 23%, 33% and 55% reduction rates mean number of items 10, 15, and 25; for SSE 2006, 8, 10 and 17, respectively.

Table 4.1 Correlations of Ability Estimates between P&P and CAT (continued)

			Threshold of SE					Fixed Item*		
Ability Estimation Method	Test Length	School Type	< 0.50	< 0.40	< 0.30	< 0.20	< 0.10	23%	33%	55%
EAP	45	State	0.876	0.908	0.930	0.951	0.961	0.909	0.938	0.961
		Anatolian	0.902	0.948	0.968	0.977	0.982	0.947	0.964	0.977
		Private	0.838	0.909	0.938	0.957	0.967	0.881	0.917	0.949
	30	State	0,899	0,958	0,971	0,977	0,980	0,907	0,949	0,979
		Anatolia	0.903	0.938	0.959	0.969	0.974	0.956	0.969	0.976
		Private	0.929	0.942	0.946	0.948	0.949	0.941	0.941	0.949

* For SSE 2005, 23%, 33% and 55% reduction rates mean number of items 10, 15, and 25; for SSE 2006, 8, 10 and 17, respectively.

To make the interpretation of the results of the simulations, plots of results for 8 different conditions (MLE/EAP, 30/45items, and Fixed Item/SE Threshold) were presented. Conditions the plots specified were given on them.

Following eight figures (Figure 4.1 to 4.8) show the correlations between ability estimations obtained from P&P and CAT formats of SSE science subtest based on different ability estimation methods, test lengths and test stopping rule. Correlations estimated for state schools were observed to be lower than those of other school types for both ability estimation methods and test termination rules.

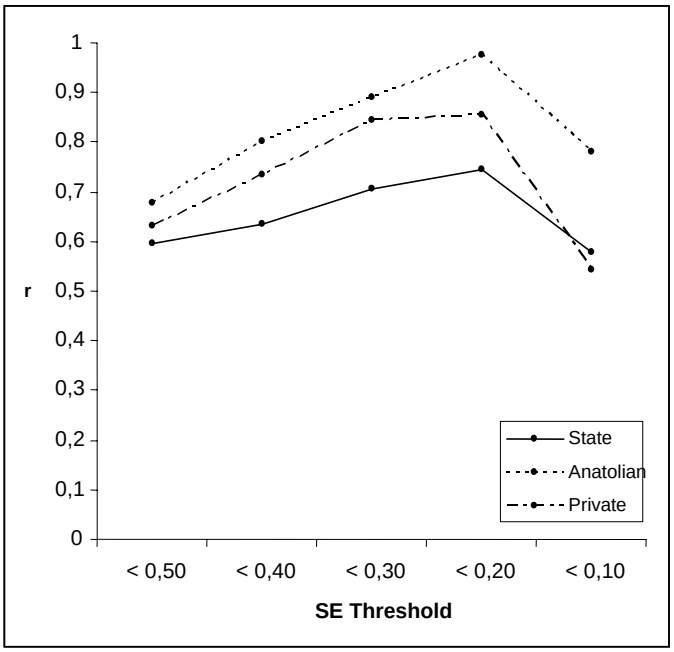

Figure 4.1 Correlations for MLE / SE Threshold / 45 items

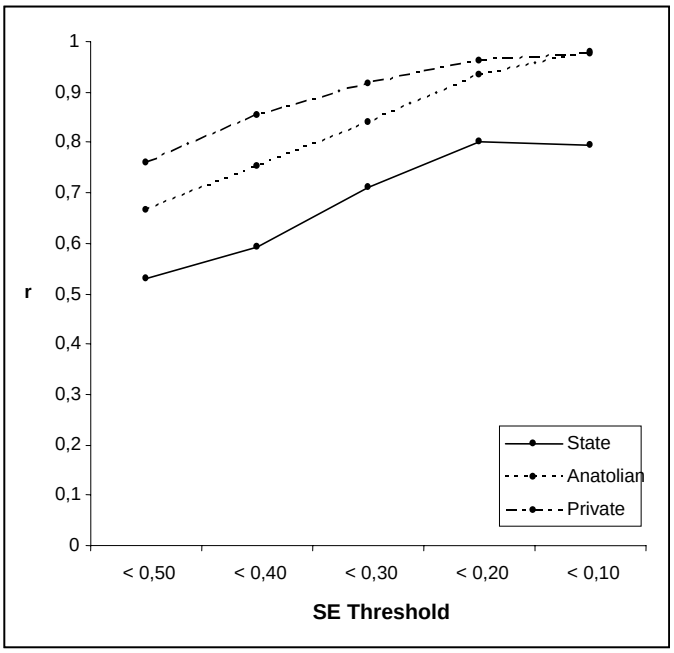

Figure 4.2 Correlations for MLE / SE Threshold / 30 Items

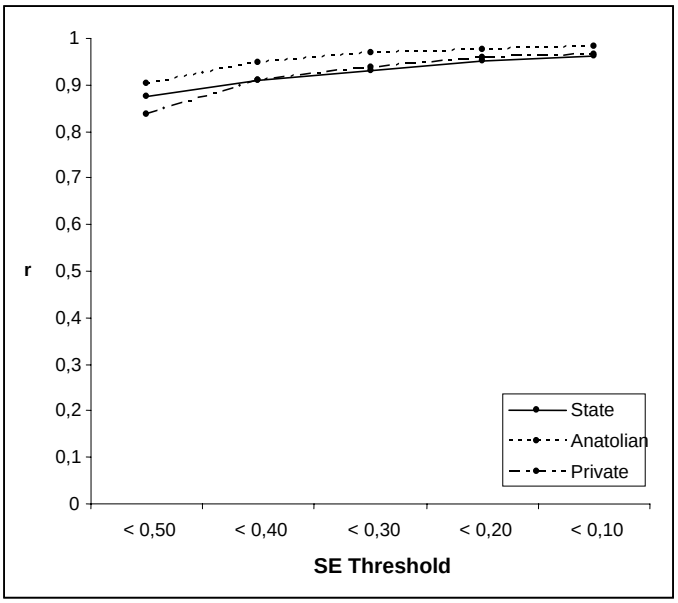

Figure 4.3 Correlations for EAP / SE Threshold / 45 items

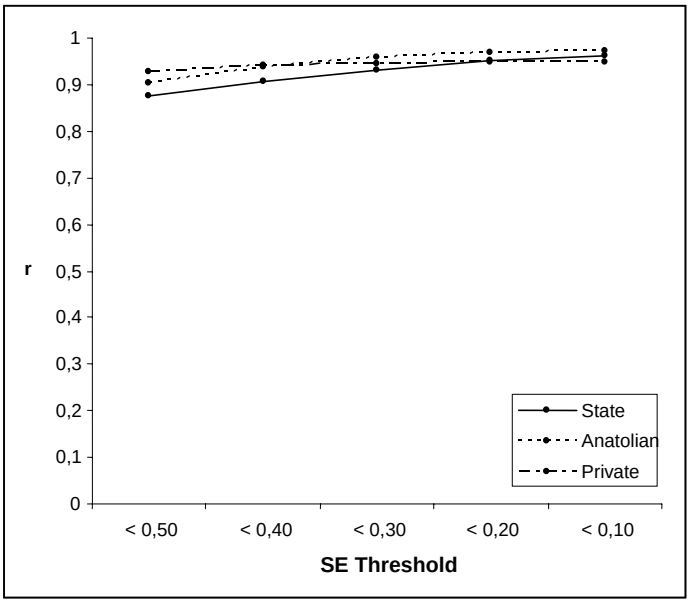

Figure 4.4 Correlations for EAP / SE Threshold / 30 items

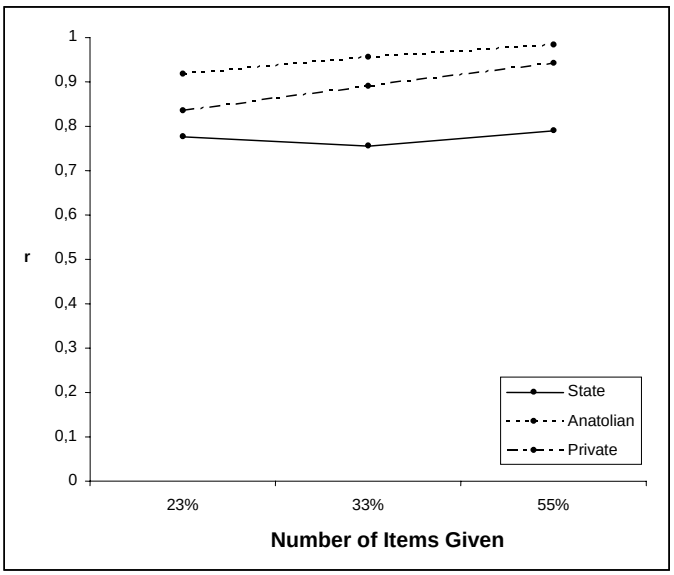

Figure 4.5 Correlations for MLE / Fixed Item / 45 items

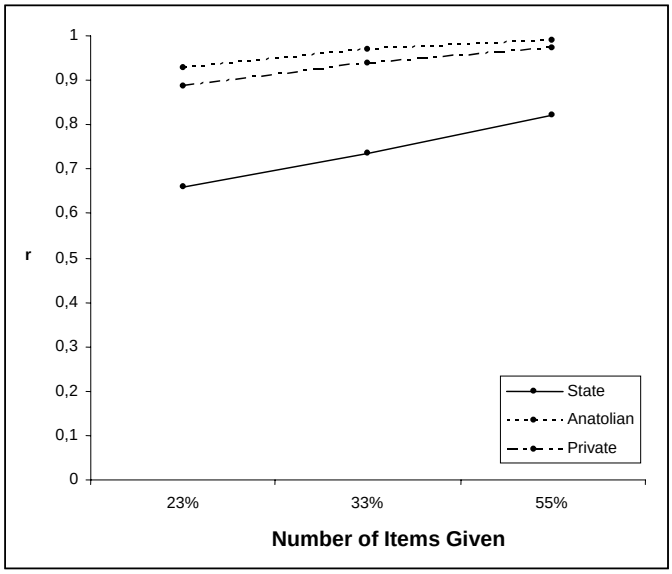

Figure 4.6 Correlations for MLE / Fixed Item / 30 items

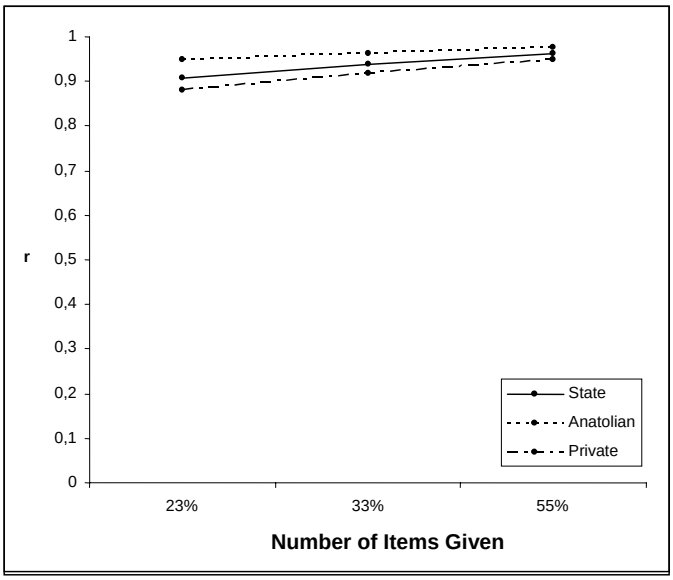

Figure 4.7 Correlations for EAP / Fixed Item / 45 items

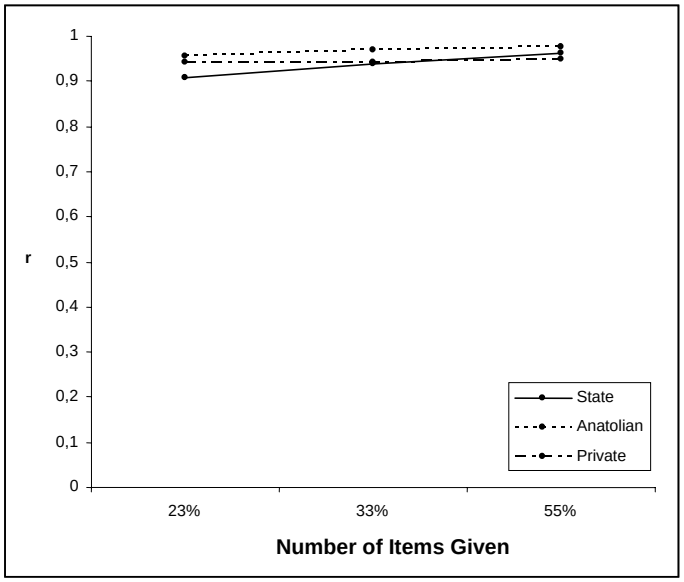

Figure 4.8 Correlations for EAP / Fixed Item / 30 items

The most striking result that can be drawn from the plots is that correlations between ability estimates based on Expected A Posteriori (EAP) method are much more larger than those obtained using Maximum Likelihood (MLE). For all conditions (different school types and test lengths), correlations between EAP estimates are higher than 0.80. And correlations between estimates for EAP are highly invariant for different conditions specified for each simulation.

As can be seen from the figures, for MLE ability estimation method, there are differences in correlations among school types. Differences are obvious for simulations results based on MLE, 45 items, and SE Threshold. Correlations obtained by simulations based on MLE / 30 items / SE Threshold seems also to be irregular when SE threshold is increasing.

Results reported up to that point supported the EAP estimation method with any test termination criteria for all conditions since correlations are 0.80 and higher. Thus, preliminary results indicated that EAP might be a better choice for ability estimation method in using for SSE.

Despite findings supporting EAP method, these results seemed not to be enough for making a statement in favor of EAP method. Findings about test stopping rules should also be investigated in performance of ability estimation methods. For example, fixed item test stopping rule may yield higher means of standard errors for the ability estimates not to be enough to obtain a reliable measurement and assessment experience. In the same way, a test stopping rule using SE thresholds could be objected if the numbers of items given in CAT are too high to make possible use of CAT rational because if the number of items approximates to those of real tests, CAT is of little use.

Therefore, to further investigate findings about performance of ability estimation method with test termination rules, in-depth analyses were conducted. For SE threshold test termination criteria reduction rates of items were investigated and SE values were compared for fixed test length (Table 4.2).

Table 4.2 Medians of Number of Items Used in Simulations for SE Threshold

Ability Estimation Method	Test Length	School Type	Threshold of SE				
			< 0.50	< 0.40	< 0.30	< 0.20	< 0.10
MLE	45	State	3	4	5	7	45
		Anatolian	4	6	8	22	45
		Private	4	6	9	23	45
	30	State	3	4	6	12	30
		Anatolian	3	4	5	8	30
		Private	4	5	6	14	30
EAP	45	State	6	8	14	25	45
		Anatolian	5	9	15	30	45
		Private	6	12	23	39	45
	30	State	6	13	19,5	30	30
		Anatolian	4	6	9	14	30
		Private	6	10	20	30	30

To make the numbers visually interpretable, following figures are presented (Figure 4.9 to 4.12).

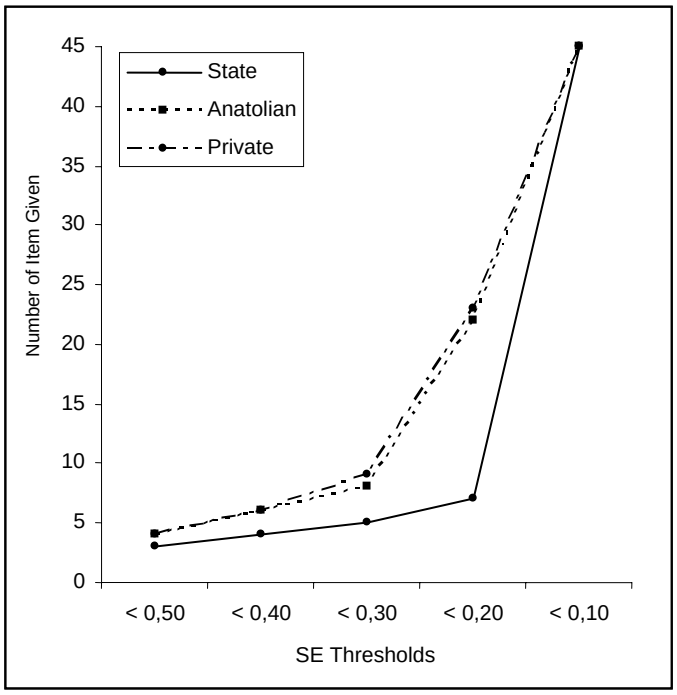

Figure 4.9 Number of Items for MLE / 45 items

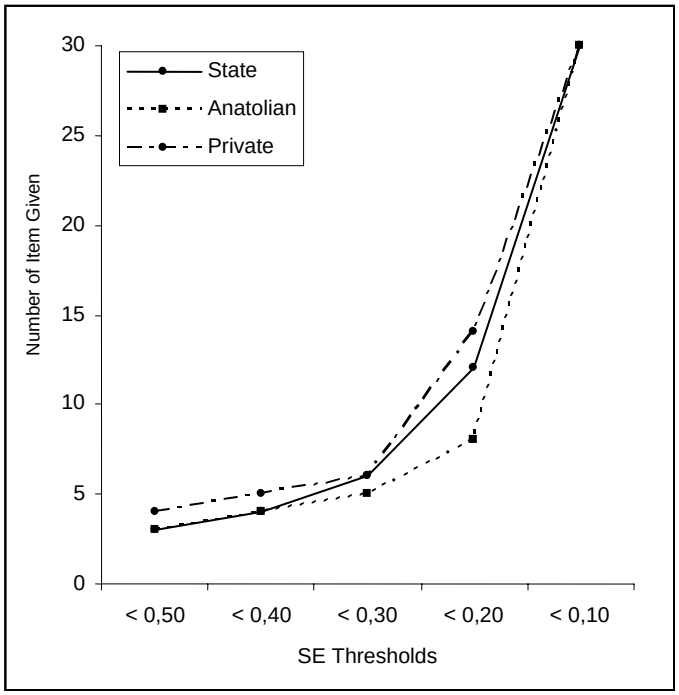

Figure 4.10 Number of Items for MLE / 30 items

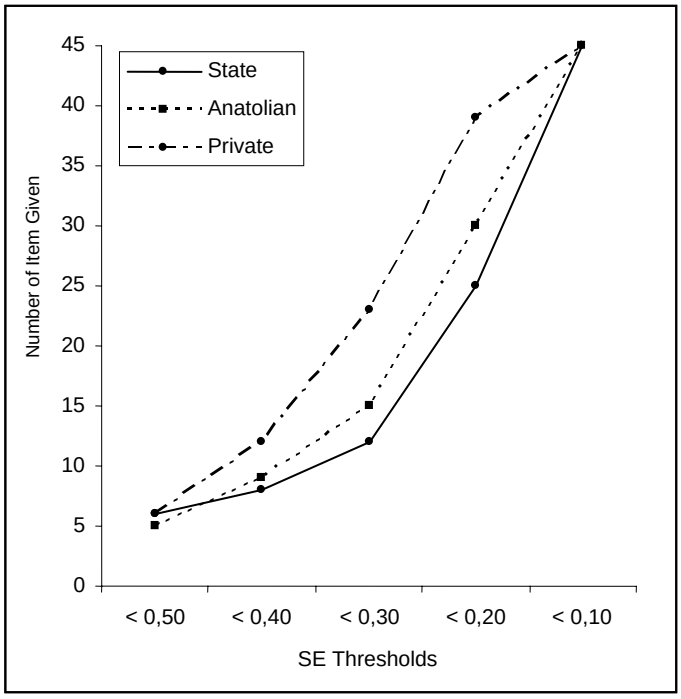

Figure 4.11 Number of Items for EAP / 45 items

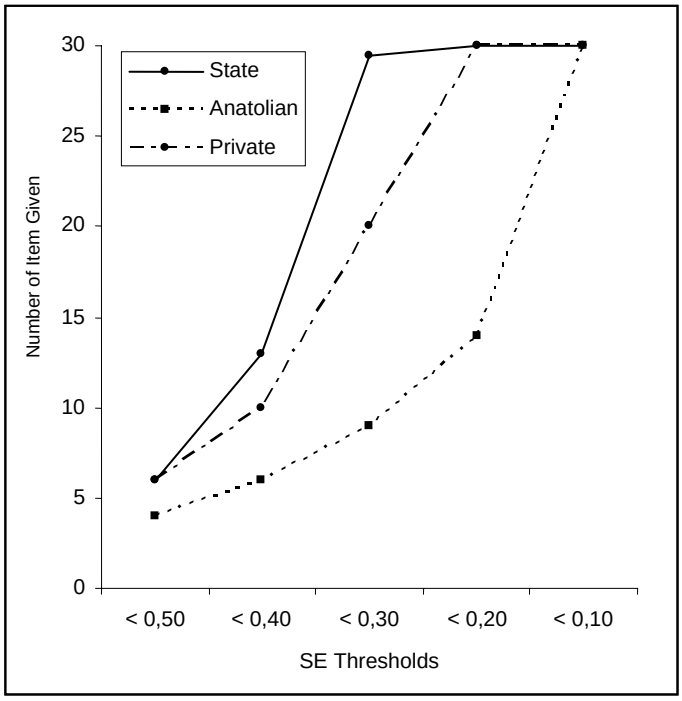

Figure 4.12 Number of Items for EAP / 30 items

Across the school types, results of simulations conducted with MLE indicated that state schools require fewer items. But for EAP-estimated post-hoc simulations, number of items required increased. For SEs of 0.10 (equal to CTT reliability of 0.99), the same number of items in the P&P SSE Science subtest demanded by CAT simulation.

For the EAP estimation, state schools of different test lengths (i.e. different years) needed different number of items to achieve to SEs of 0.30 and lower. Post-hoc simulations with a test length of 45 (SSE 2005) needed fewer item to obtain 0.30 SE, on the other hand state school were given 98.33% of the item bank that includes all items of SSE 2006 Science subtest (30 items). Anatolian high school which accepts students using a selection test required less number of items, like private schools.

As expected, number of items required to higher degree of standard errors is much more than those of lower SEs. To reach a SE of 0.10 numbers of items are the same with P&P full test length. All simulation results indicate that to achieve 0.10 degree of SE there can be any reduction in the number of items. Especially for EAP estimation method, number of items used in CAT simulations differentiated.

On the other hand, a SE of 0.30 (equal to as CTT reliability 0.91) can be achieved using reasonable number of items. For example, mean of items required to obtain a SE of 0.30 is 6.5 for MLE, on the other hand EAP needs a mean of 18.42 to achieve to the same SE level. Reduction rates for each simulation situation are given in Table 4.3.

Although MLE seemed to be a better choice, number of items required by EAP to make ability estimates with SEs of 0.30 is also not too high and EAP can still be a choice. Also number of items demanded by MLE can be harmful for content validity of the test. Another problem that may be stem is the increasing chance of blind guessing. Probability of blind guessing for items with 5 alternatives in SSE science test is $(1/5)^5 = 1 / 3125 \sim 0.03\%$, which is too low for SSE taken by hundreds of thousands students. For this reason EAP may be good alternative for its higher correlations and reasonable reduction rates in the number of items given to examinees.

In Table 4.3 it can be seen that reduction rates for CATs with MLE are better than those with EAP. Since it is well known that EAP estimation method requires more item than MLE to attain the same standard error, lower reduction rates are expected for EAP. But both ability estimation methods, MLE and EAP, indicate that significant reduction rates can be obtained for CAT administration using a test termination criteria of SE of 0.30.

If correlations are investigated according to the school types, when MLE was used to obtain a SE of 0.30 or higher reduction rates of approximately 16% to 50% was achieved. On the other hand, EAP reduction rates for the same SE levels with MLE.

After investigating reduction rates for fixed SE test termination rule, in a similar manner fixed test length method is investigated in SE of ability estimates (Table 4.4).

For SSE 2005, 23%, 33% and 55% reduction rates mean number of items 10, 15, and 25; for SSE 2006, 8, 10 and 17, respectively.

Table 4.3 Item Numbers Used in Percentages of P&P SSE in Post-hoc Simulations

Ability Estimation Method	Test Length	School Type	Threshold of SE				
			< 0.50	< 0.40	< 0.30	< 0.20	< 0.10
MLE	45	State	6.67	8.89	11.11	15.56	100.00
		Anatolian	8.89	13.33	17.78	48.89	100.00
		Private	8.89	13.33	20.00	51.11	100.00
	30	State	10.00	13.33	20.00	40.00	100.00
		Anatolian	10.00	13.33	16.67	26.67	100.00
		Private	13.33	16.67	20.00	46.67	100.00
EAP	45	State	13.33	17.78	31.11	55.56	100.00
		Anatolian	11.11	20.00	33.33	66.67	100.00
		Private	13.33	26.67	51.11	86.67	100.00
	30	State	20.00	43.33	98.33	100.00	100.00
		Anatolian	13.33	20.00	30.00	46.67	100.00
		Private	20.00	33.33	66.67	100.00	100.00

Table 4.4 Median of SE Values for Different Test Lengths

Ability Estimation Method	Test Length	School Type	Fixed Item		
			23%	33%	55%
MLE	45	State	0.16	0.14	0.12
		Anatolian	0.26	0.23	0.19
		Private	0.28	0.24	0.19
	30	State	0.21	0.18	0.17
		Anatolian	0.17	0.15	0.13
		Private	0.23	0.19	0.17
EAP	45	State	0.43	0.36	0.26
		Anatolian	0.39	0.34	0.30
		Private	0.44	0.40	0.34
	30	State	0.45	0.44	0.42
		Anatolian	0.33	0.26	0.18
		Private	0.43	0.40	0.37

Figure 4.13 to 16 presents medians of SE values if all examines are given the same number of items for CAT simulations, that is, fixed test length test termination rules is used. As knows, EAP method produces more SE than MLE for the same response pattern. Values of SEs for EAP estimation method are larger than those of MLE. As expected and seen in Figures 13 to 16, there is a declining trend line with the number of reduction decreasing. That is also normal, since while reduction rate decreases, i.e. number of items given increases, SE values are contributed by more items, producing higher SE values. This is due to mathematical formulation of EAP. Ability estimations of state schools simulated with EAP with 30 items yielded more SEs than SSE simulation of test length of 45.

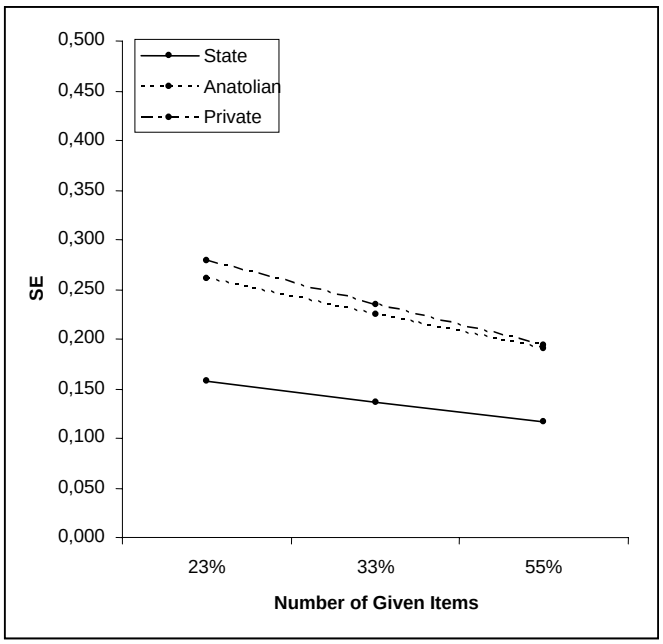

Figure 4.13 SE Levels for MLE / 45 items

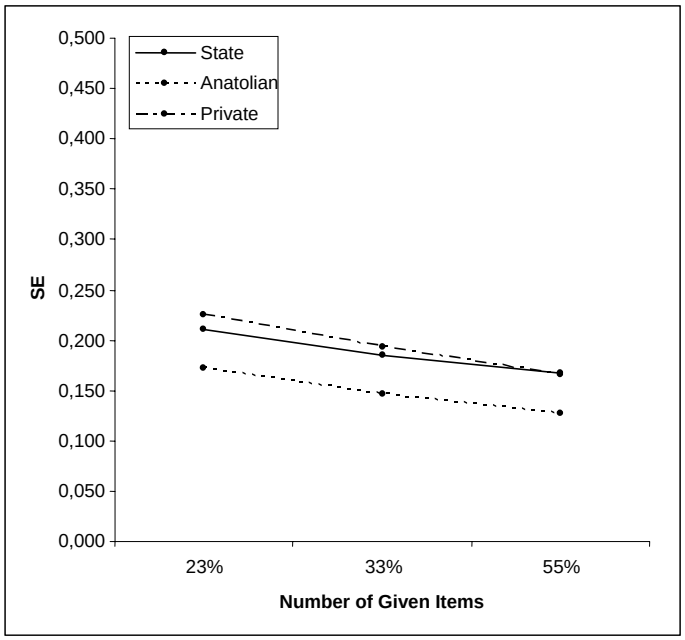

Figure 4.14 SE Levels for MLE / 30 items

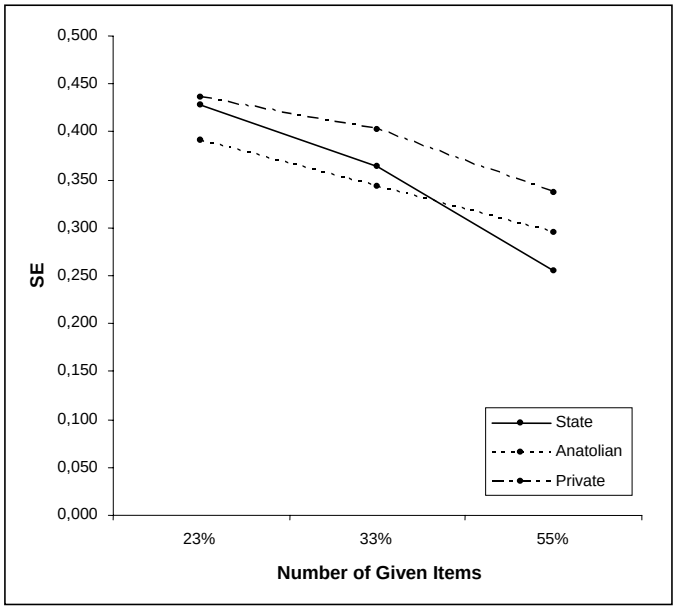

Figure 4.15 SE Levels for EAP / 45 items

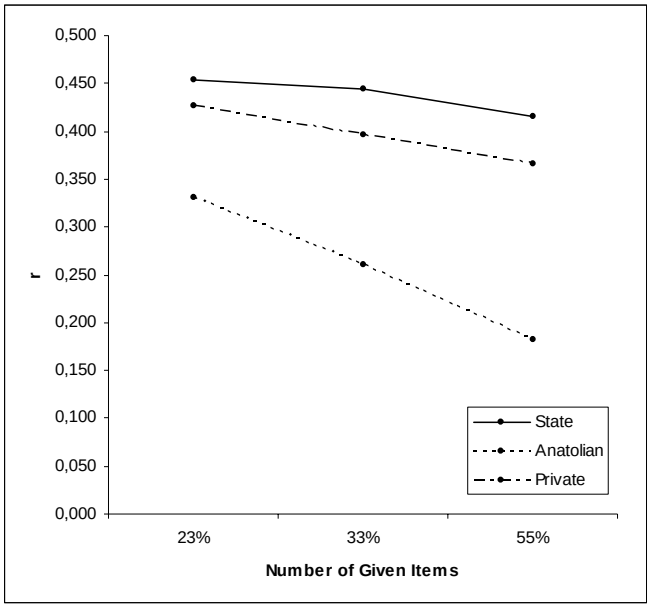

Figure 4.16 SE Levels for EAP / 30 items

Investigation of SE levels for different test lengths indicated that MLE produced lower SEs, and thus required fewer items, as expected. On the other hand, EAP produced higher correlations with real ability estimates of examinees from real SSE. EAP generates higher SEs, requiring more items; however number of items used by EAP is moderate for being considered for a choice for ability estimation for SSE science subtest. EAP overcome the main disadvantage of MLE, diverging ability estimation arising when examinees provided zero or full response patterns. This is not a uncommon situation for SSE science subtest, many students response a few items with zero correct, especially by students of state schools. EAP is capable of discriminating ability levels of examinees who provided zero correct response patterns, which is a point that MLE do not achieve. Also for examinees with highest ability examinees, such as students from science high schools, all-correct response pattern is another diverging ability estimation problem encountered. Table 4.5 presents the percentages of the examinees whose

ability estimations could not be produced using MLE with both fixed SE and fixed test length termination rules.

Table 4.5 Percentage of Unestimated Examinees for MLE

Test Length	School Types	Fixed SE					Fixed Test Length		
		0.5	0.4	0.3	0.2	0.1	23	33	55
45	State	33.72	33.42	33.37	33.70	33.88	33.87	33.93	34.02
	Anatolian	24.40	24.13	23.85	24.18	23.80	24.18	23.98	0.00
	Private	16.73	16.87	16.63	16.70	17.80	16.50	16.73	16.60
30	State	40.20	39.98	40.13	40.73	40.95	40.48	40.83	40.30
	Anatolian	21.73	21.77	21.97	21.57	21.33	21.50	21.03	21.90
	Private	30.33	30.30	29.90	30.28	29.83	30.70	30.10	30.05

Although there is no differences between rates in terms of test termination criteria, there are differences between school types and test lengths. State schools have the highest missing rates for both test lengths. On the other hand, the other school types have lower rates of unestimated examinees. The Table 4.4 provided findings in favor of EAP since as can be seen from the table diverging test termination is a problem in estimating examinee abilities using MLE.

An additional analysis conducted to how many examinees were left at an unacceptable SEs estimation with fixed test length test termination rule. With EAP fixed test length produced a group of examinees changing in sizes of 10% to 20% had SE estimates over 5.4, which is a value that corresponding to CTT reliability below 0.70. Also regarding the test taking behaviors of different examinee groups (such as state schools), fixed test length may cause to some problems. If an examinee produces an aberrant test behavior (blind guessing, correct to very hard/wrong to very easy items, etc.) cause a delay for obtaining a acceptable SE. For some examinees even all items are given, SE estimate can not obtained to indicate a reliable measurement experience.

Based on these reasons, EAP ability estimation and a SE threshold (fixed SE) were selected for ability estimation method and test termination criteria, respectively for a CAT administration for SSE Science subtest to provide reliable and consistent ability estimations with P&P SSE sessions.

4.2 Live CAT Administration

To investigate how CAT works for a group of examinees, a live CAT administration was conducted to real examinees took SSE science subtest six months before than the time of CAT administration. Ability estimations of participants obtained from CAT and P&P formats of SSE science subtests were compared.

Thirty three examinees were participated to CAT administration that were set up with EAP estimation method and fixed SE test termination rule. Threshold for SE was taken to be 0.30 to provide a highly reliable testing session.

A high proportion of the live CAT administration is from Anatolian high schools. That kind of schools uses a selection procedure to accept students. Therefore student profile of Anatolian high schools include student that can be located in higher levels of ability continuum. Nearly half of the participants (14 examinees with a proportion of 42.4) did not take a computer-related course.

All participants received P&P SSE science subtest in year 2007. Median of correct responses is 28 out of 30 which corresponds to 93.33% of items. Medians of ability and SE estimations were found to be 1.21 and 0.22, respectively.

Correlations between ability estimations obtained from CAT and P&P SSE Science subtest was found to be 0.736 ($p<0.05$). Figure 4.17 presents scattergram of correlations.

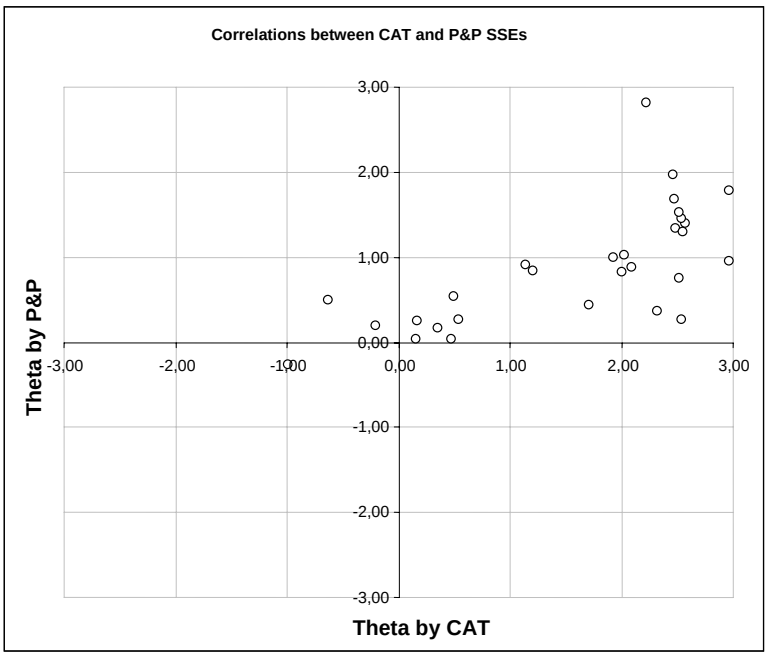

Figure 4.17 Relationship between CAT and P&P SSE Science Subtest

When numbers of items given to examinees were investigated, it was found that five examinees with ability levels 0.53, 0.50, 0.25, 0.26, and 0.26, respectively were given only 4 items with SE values below 0.30 for all.

Two examinees with ability levels of 2.82 and 2.66 were given 30 items (i.e. whole items in P&P SSE 2007). SE values for these two examinees were found to be 0.382 and 0.373.

Median of number of the items given to examinees in live CAT administration phase found to be 9.0 indicated a reduction rate of 0.70 of P&P science subtest.

Ability estimations of twenty-seven participants were observed to be lower than those of P&P. Only six examinees had ability estimations greater than in CAT administration. Medians of SE estimates of CAT are less than those of P&P SSE. Eight examinees' SE estimations from CAT administration were found to be larger than P&P. Twenty-eight examinees received to 0.30 SE thresholds, obtaining highly reliable ability estimates. For two of the participants SE threshold could not be reached. However these two participants had SE values of 0.373 and 0.382, respectively (Table 4.6).

Table 4.6 Ability Estimations of Live CAT and P&P

	P&P			CAT		
	Ability	SE	# of Items	Ability	SE	# of Items
Mean	1.77	0.47		0.99	0.28	12.36
Median	2.22	0.50		0.91	0.29	9.00
Sd	1.22	0.22	30	0.76	0.03	7.83
Minimum	-0.99	0.17		-0.27	0.23	4.00
Maximum	3.55	0.99		2.82	0.38	30.00

As can be seen from Figure 4.18 and 4.19, ability estimations of 27 examinees were found to be lower compared to P&P. In the same way SE values of ability estimations of 25 examinees had lower SE values than those obtained from P&P SSE science subtest. All examinees' SE values were below 0.30.

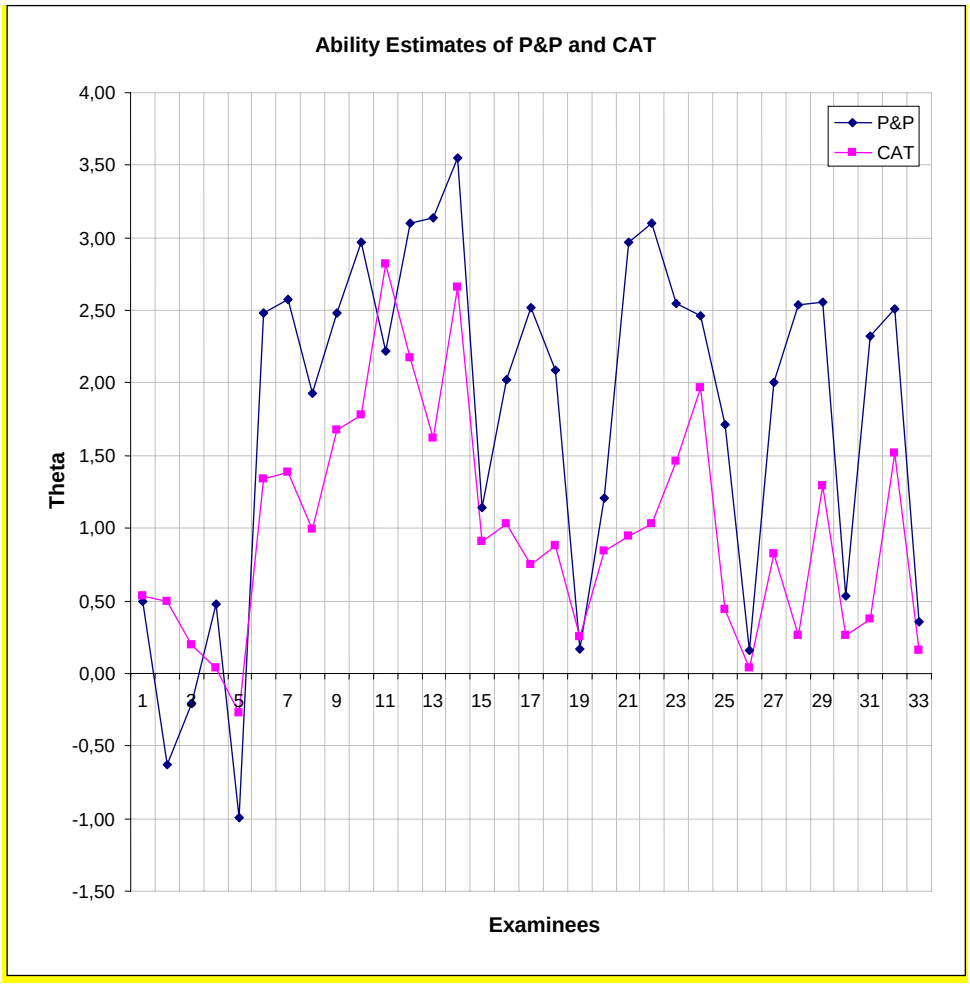

Figure 4.18 Ability Estimates from CAT and P&P SSE Science Subtest

This is due to psychometric properties of SSE. In P&P SSE examinees are given all items, fitting or not fitting their ability levels. On the other hand, in CAT

administration examinees are encountered with only items tailored to their ability levels. In other words, in a tailor test, examinees do not see items with extreme difficulty compared to their ability and. in turn IIFs are summed using item with maximum information for each examinee.

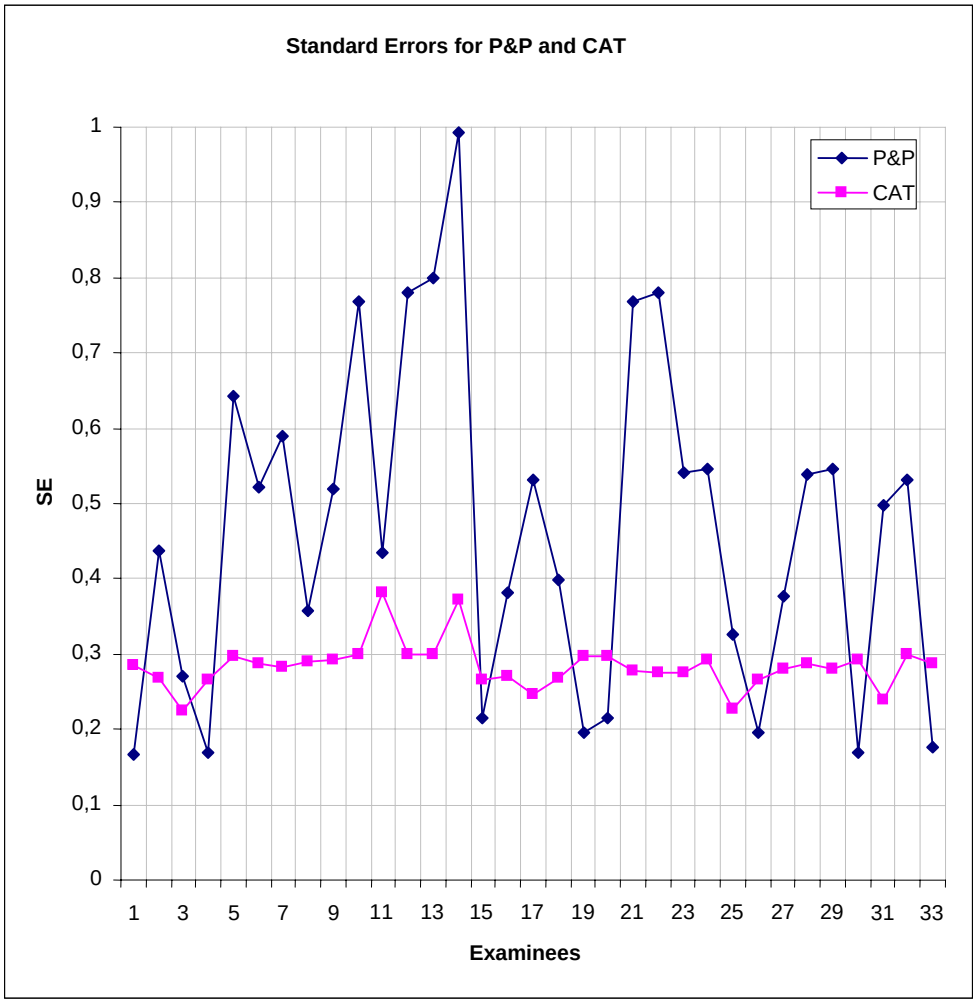

Figure 4.19 SE Estimates from CAT and P&P SSE Science Subtest

Results pointed out that CAT administration used fewer items, produced lower SE estimates and thus provided reliable testing sessions. Figure 4.18 and 4.19 presents ability and SE estimates both from CAT and P&P administrations.

CHAPTER 5

CONCLUSION AND DISCUSSION

In this chapter, results of the present study are summarized and discussed. Results of the present study can be divided into two groups: simulations and live testing. First results obtained from the post-hoc simulation and live CAT testing phases are discussed. Then issues related to applicability of CAT format to SSE are discussed. Lastly, limitations of the study are given and lastly suggestions for future researches are given.

The purpose of the present study is to compare results of SSE science subtest obtained from CAT and P&P formats considering ability estimation methods and test termination rules.

Different high school types (state, Anatolian, and private) and different test lengths (45 items for SSE 2005 and 30 items for SSE 2006) were included in the study to investigate performance of CAT administration on different ability and test characteristics.

First using post-hoc simulations effects of CAT strategies on recovery of ability estimations of real examinees were examined. After simulation phase, using real examinees a live CAT administration was conducted to observe the performance of CAT in a realistic situation.

5.1 Summary of the Findings

- Ability estimation using MLE yielded lower correlations for all test lengths and school types.
- EAP estimation method produced ability estimates which are highly correlated with P&P SSE science subtest than MLE did.

- Fixed test length test termination rule had a better performance in terms of correlations between ability estimates of CAT and P&P SSE science subtest.
- There are no differences between different test lengths in ability recovery for a given ability estimation method.
- For MLE ability estimation method, even tough correlations between ability estimates increases with SE levels decreasing, a SE level of 0.10 could not be well estimated.
- For MLE there are differences in terms of ability estimation across school types.
- In EAP school types were not observed as a factor differentiating ability estimations.
- MLE estimation method needed fewer items than EAP to estimate examinees' abilities. As expected number of items required increased with SE level decreasing.
- EAP method used more items than MLE did for ability estimation.
- There are no differences between correlation in terms of test lengths and school types except higher SE levels.
- A SE level of 0.10 could not be achieved without giving all items of SSE to examinees.
- For fixed test length MLE produces lesser SE values to estimate ability. EAP, due to its nature, yielded higher SE estimates.
- MLE ability estimation method left a group of examinees without estimating their ability levels. Especially for state high schools size of group left unestimated is higher.
- Live CAT administration yielded a strong relationship between ability estimates of examinees from CAT and P&P administration of SSE science subtest.
- Ability estimates of CAT were lower than those of P&P.

- SE values of CAT ability estimations were lower than those of P&P SSE.

5.2 Post-Hoc Simulation Phase

The post-hoc simulation phase was conducted to investigate the performance of CAT strategies on ability recovery.

Post-hoc simulation studies yielded results supporting the applicability of CAT administration in SSE science subtest. One of the main advantages of CAT that can be stated as higher reliability with fewer items seemed to be confirmed in the context of the present study.

Ability estimation methods covered in the present study, MLE (Birnbaum, 1958) and EAP (Bock & Aitken, 1981), worked in parallel to literature. As Hambleton and Swaminathan (1984) stated MLE did not work in some situations in which examinees provided full or blank response patterns. On the other hand, MLE estimation method was able to produce less SE error with the same response pattern compared to EAP. Thus MLE method could be a good choice, however for SSE science subtest nondivergent estimates could easily not be met as stated in Chapter 4 Results. Ratios of examinees left unestimated using MLE due to full wrong response pattern can be go up to 40%. For these circumstances a Bayesian approach may be helpful since Bayesian ability estimation methods are capable to yield ability estimates in all cases (Hambleton, Swaminathan, and Rogers, 1991). It is important to note that ability estimates produced by EAP are higher than those of MLE, but higher SE values are preferable over left-unestimated abilities. State schools have the highest unestimated groups of examinees for both test length.

Also for state schools representing the largest examinee group taking SSE science subtest, correlations between CAT and P&P obtained using MLE observed to be lower than the other groups. This can be explained by psychometric properties of examinees in state school group such as random guessing, low cognitive levels, etc. and MLE estimation method. Examinees from

state school may show interesting test behaviors such as aberrant response pattern. For example they provide wrong response to very easy items or true response to very hard items solely by blind guessing or eliminating some of the alternative and choosing one from the rest. MLE method uses fewer items, providing less SE value for ability estimate, however it is not capable to eliminate or reduce the effect of such unexpected behaviors from examinees.

In the present study, EAP method produced higher correlation between CAT and P&P for all test lengths and school types with test termination rules were hold constant. Since EAP assumes a prior distribution about examinee ability (Bock and Mislevy, 1982), it is capable to produce strong ability estimates than MLE does.

Thus it can be stated that for examinee groups such as those who take SSE science subtest EAP estimation method should be used since one true/one false response patterns required by MLE can not be met by many examinees (i. e. up to a level 40%). Despite of advantages of MLE provides, it did not work well for examinee groups in SSE science subtest, especially for state school examinee groups.

As to test termination rules, two approaches were selected for the present study: fixed test length and fixed SE. With ability estimation method fixed, no significant differences between ability estimations observed for both different test length and school types.

In the literature, fixed SE test stopping rules was favored (Lord & Stocking, 1987; Babcock & Weissm 2000), because it guarantees for each examinee a reliable ability estimation is made. On the other hand, when using fixed SE test termination rule number of items required to estimate ability is important. A reliable ability estimation providing no reduction in the number of items may not be preferred. However fixed test length provides a limitation to avoid examinees taken too many items, but this time after giving all items examinees ability levels can not be estimated with desired reliability level. For the present study, a combined test termination approach was adopted. Main test

termination rules is fixed SE but for an examinees who took items in the same number of P&P without obtaining a reliable estimate test is stopped.

For fixed SE method as SE level got lower, that is, reliability got higher, number of items required increased. For a 0.10 SE level, no reduction in the number of items was observed. For EAP SE level of 0.20 was also hard to satisfy, while MLE was more liberal, using fewer items. A SE level of 0.10 is too high to be expected to obtain in terms of item numbers, therefore it is not so important not to achieve that SE levels. A SE level of 0.30 would be enough for ability estimation, corresponding to classical reliability of 0.91. For EAP ability estimation method, 30-item test showed lower reduction rates than MLE did. This is due to post-hoc simulation used. Simulation phase of the present study conducted a small CAT administration with items only in respective SSE. That is, for simulation of 2006 only 30 items were used for CAT simulation. Therefore it was difficult to find items for each examinee to get lower SE values and different reduction rates could be observed.

For fixed test length, SE values were investigated for school types and test length. EAP / MLE difference in terms of SE is confirmed again by the results. EAP produced higher SE values. Fixed test length approach was also investigated in terms of SE values in combined with EAP, since ability EAP was selected as ability estimation method. For some high school types and test lengths median of SE of ability estimations go up to 0.45. That indicated that fixed test length could provide higher SE values, that is, unreliable ability estimates.

Between two test termination rules fixed SE level seemed to be a better choice. Although it needs higher number of items required, fixed test length may provide unreliable ability estimates, which may cause to a more serious problem compared to the former.

In general different findings of the study indicated different results. On the other hand, findings should be investigated in a combined way. For example, at first MLE seemed to a better estimation method but when unestimated examinee ratios were investigated, EAP seemed to work much better. In a similar way, the

largest reduction rates obtained with state high schools, however correlations with real SSE science test indicated weak relationships.

As a result of the post-hoc simulation phase, no differences observed between school types and test lengths especially for EAP / 0.30 SE CAT testing strategy, indicating that CAT can be applied to SSE with different school types, i.e. different ability groups.

After selecting testing strategies for CAT through simulations, these strategies were used with live examinees in CAT session.

5.3 Live CAT Administration Phase

Post-hoc simulation phase included small simulated CAT sessions using 45 or 30 items and that limited the results since a real CAT needed a large item bank to select items from. Also administering a live CAT to real examinees may reveal the effects related to CAT format. To this end, a real CAT session was included in the present study.

Live CAT session was set to have a SE level of 0.30 and EAP ability estimation method. 33 participants took CAT format of SSE science subtest from an item bank including items from older SSE science tests.

Correlation between ability estimations of CAT and P&P SSEs were found to be 0.736, a value high enough to be supporting evidence of applicability for CAT in SSE science subtest and also for other subtest and large-scale testing programs. Correlation could lessen due to some reasons. Students in the sample live CAT phase were given items that the students saw them before when they prepared for higher education entrance examination. Also items given to examinees are mainly developed for moderate ability groups therefore computer algorithm might have difficulty in finding proper items for participants and ability estimations produced can be biased.

The correlation found was lower than those obtained in post-hoc simulations. This is an expected situation, since many factors may intervene the live CAT phase. Factors such as computer anxiety, not taking CAT session

serious unlike real SSE, requirement of using computers for test, also for prohibition of omitting and moving behavior could bring explanations to lower correlations (Hambleton & Swaminathan, 1984; Lord &Stocking, 1968; Sands, Waters &McBride, 1997; Rudner, 1998; Cikrikci-Demirtasli, 1999).

Since participants of live CAT administration took 2007 SSE science subtest, additional test stopping rules applied in post-hoc simulation was also used here. If an examinee was given 30 items and still had a SE level of above 0.30, testing session was stopped. This rule applied for two high-able participants, for whom test sessions were stopped after 30 items were given. These two participants have 28 and 30 correct responses and ability estimations of 2.22 and 3.55 in 2007 P&P SSE science test. For these participants SE levels could not be obtained since item bank of CAT sessions included only items from older SSE science subtests. Since SSE difficulty levels of science items had a mean of 0.889, item information functions could not find appropriate items for these two participants. When algorithm could not find appropriate items, it selected the nearest proper items which are not perfectly fit to their ability levels.

The fact that CAT provides a reduction in the number of items required producing less SE, that is, higher reliabilities (Embretson, 1996) once again confirmed by live CAT phase, for 25 participants' SE values was observed to be lower than their P&P SE values. On other hand SE values of remaining 8 examinees were higher compared to P&P, however they were still below 0.30.

The number of items given to participants provided a significant reduction rate. Mean of the items given to examinees is 12.364. This result, combined with SE values all below 0.30, supported strongly use of CAT for SSE science subtest.

Out of 33 participants, 27 have CAT ability estimations lower than P&P estimations. For 6 of them, CAT estimations were found to be higher.

In CAT items selection is conducted using related algorithms. By this way, item difficulty and ability of examinees were tried to be kept in parallel as much as possible and as a result of that, examinees usually were given optimal items in terms of difficulty and not encountered too easy for their ability levels. In other

words, CAT gives harder items to examinees and this cause to lower (but more reliable) ability estimations.

Another interesting results obtained from the live CAT administration was that nearly half of the participant (42.4%) did not take a computer-related course. Despite that, as findings indicated, all examinees were able to use computers without any problem and attend to CAT administration. Thus application of CAT format seemed not to be affected from familiarity a formal course related to computers.

Live CAT administration provided promising results related to CAT administration for SSE science subtest.

5.4 Applicability of CAT for SSE

Results of the present study indicated that SSE could be administered using CAT format. Using CAT format of SSE science subtest not only produced highly correlated results with P&P SSE scores, also more reliable ability estimates with fewer items.

It is important to note that the present study investigated applicability of CAT administration from the perspective of measurement and assessment. To convert P&P SSE to CAT SSE, more research should also be conducted on other fields than measurement such as logistic, computer opportunities, test and design computer application interfaces. Numbers of computers at testing centers and network security of item bank are points that should be studied. Also anxiety of using computer is a point that should be investigated.

In measurement perspective, item bank for CAT SSE should be developed, including items that cover a broad range of difficulty so that computer could find items having proper difficulty for each examinee.

Using an item bank not fitting ability level of examinees results in over-exaggerated item weights just is the current situation in SSE. Since items given to examinees in P&P testing format do not match to ability levels, means of total scores of SSE are so low in all subtest that only giving a true response to an item

response changes examinees orderings significantly. If correct answer is provided by blind-guessing an examinee's score can also be much higher than deserved just by chance. The probability that providing correct responses to 5 items in a 5-alternative test is $1/2^5 \approx 3\%$. This value is not so low for a test taken by over one million examinees. On the other hand, if ability-difficulty level matching can be achieved examinees encounter only with proper items without dealing with especially too much items. CAT is a systematic methodology capable to achieve that aim. Also pseudo-guessing parameters estimated for the items in the present study indicated that for some items chance factor included. Blind-guessing or somewhat smart blind-guessing which can be done eliminating some alternatives for correct response seemed to be a factor that affect examinees ordering in the present study.

Item bank should also large enough to avoid reveal of all items in a short period of time. Remembering items or encountering with old items is certainly threads to validity of test. Student Selection and Placement Center have administered examinations for entrance to higher education programs for long years and have a large item pool. With minor modifications for example on item contents produce a large item pool that can be used in CAT administration.

Also CAT administration provides test developers with an opportunity to use new item format that are not possible in P&P tests. Items with animations, audio-video records, user interaction, etc. can effectively be used.

Test scoring and security of test documents are other problematic areas in P&P test administration. CAT administration applied in the present study produced examinee scores immediately after test termination. Also since there is no need to use paper-based answer sheets, transportation and security issues were eliminated.

Cheating or collusion detection analysis is also possible in CAT as in P&P. Statistical tests using examinees' response times to items are developed by researchers (Wise & Kong, 2005; van der Linden, 2008).

Public relations are a hugely important issue that can be addressed by explaining the principles of CAT administration. Since examinees tested by CAT administration encounter with different items fitting to their ability levels, this may cause of a serious concern among public. People may criticize CAT format from differentiating items for each examinee, equality of test scores obtained from different items for the same SSE and this also results in controversies. These issues should be explained carefully by those related to the public.

Application of CAT format for SSE may be accompanied by a change in test giving conditions. If issues such as testing centers, Internet connection, network security, etc. are well established. CAT format of SSE may be suggested to be given three times for each examinee in, to say, summer session. Test scoring is instantly made therefore there may be enough time to test examinees again if necessary. The highest scores of examinees may be used to selection and placement to higher education programs. By this way, one of the major criticism against SSE that can be stated as a using 3 hour-test is too critical, examinees who have problems at the test date loose their chances, and have to want one year for the next, test can be overcame.

SSE is the major central large-scale testing administration project. Over 1 million examinees take the SEE each year. Transforming administration type of SSE may be detrimental. Therefore CAT format of SSE can be optional, explaining the both formats' similarities and differences.

Another solution can be using CAT format in another large-scale testing programs to make examinees and public gain familiarity with CAT administration. For example, the Entrance Examination for Graduate Studies is given by university students and graduates for graduate studies. Population of that examination is from higher levels in educational level and university graduates are more familiar with computers. Thus application of that examination prior to SSE can be helpful in both examinees attitudes toward CAT format and public concerns.

It is expected that the present study constitute a basis for investigation of applicability of CAT format to SSE.

5.5 Limitation of the Present Study

One of the limitations of the present study is psychometric weakness of the items of SSE. Especially for state high school students item difficulties are high above of the average ability level of students. This is a major problem, not only for CAT administration, but also for P&P format of SSE subtests.

For CAT an item bank not including items with a broad range of ability of examinees may produce an item selection problem because computer algorithms developed to select items fitting difficulties to ability levels can not find appropriate items for examinees outside the difficulty range of items. Items used in post-hoc simulations and live CAT testing do not exactly match to examinees in terms of ability.

Another limitation is about participant of live CAT administration. Since participants are highly-able groups compared to examinees taken SSE, generalizability of the results of the present study can be a problem of objection.

Also item bank of live CAT phase includes items of older SSEs, item selection is again a problem in finding items with proper difficulty. In the present study, item selection procedure (Maximum Information) used a narrower item bank to find items providing maximum information, i.e. matching ability levels of examinees in difficulty.

Item exposure may be another dimension that can be studied. If proportions of items from different subdomains are controlled, that can provide a control over content validity of the tests.

The present study included a uni-dimensional science test for CAT. It is also possible to develop multidimensional CAT administrations with special emphasis on composite scores estimated through a multidimensional trait.

5.6 Suggestions for Further Research

In the present study, 3PL logistic model was used to calibrate the items both for post-hoc and live CAT phases. On other hand, 2PL model can be used for calibration so that pseudo-guessing parameters may be excluded. Also Nominal Response Model (NRM) (Bock. 1972) can also be used for calibration and scoring the examinees. NRM allows using response patterns in the form of alternatives rather than converting responses to dichotomous item response patterns.

Although the present study is about CAT administration of SSE science subtest, other subtests (mathematics, social sciences and Turkish) can also be investigated using the same research design. Also applicability of other large-scale testing administrations such as the Foreign Language Examination for Civil Servants, the Entrance Examination for Graduate Studies for CAT format can also be investigated.

Item bank of the present study includes 242 science items extracted from older SSE science subtests. The present study can be replicated using a larger item bank. Also new items can be written other than older items of SSEs and calibrated to obtain a larger ability range.

A larger and broader groups of examinees in ability range may be invited to participate to live CAT administration sessions.

Other CAT strategies not covered in the present study can be investigated. For example, item selection rules are not in the scope of the present study. There is a single item selection rule (Maximum Information) used. The other item selection procedures can be studied to observe the effect of them on recovery of examinees. Also item skipping and moving along the items are not allowed in the present study, however investigation of these CAT strategies can be helpful in providing evidences as to effect of tem ability estimation of examinees.

REFERENCES

Akyildiz, M. (2003). Klasik test kuramina ve 3 parametreli lojistik modele gore hesaplanan yetenek olculerinin oss puanlari ile iliskisi. Ankara University.

Ankenmann, R. (1994). *Goodness of fit and ability estimation in the graded response model*. Unpublished manuscript.

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.

Ben-Porath, Y. S., Slutske, W. S., Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the mmpi. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*(1), 18-22.

Berberoglu, G. (1988). *Seçme amacıyla kullanılan testlerde rasch modelinin katkıları*. Unpublished doctoral dissertation, Hacettepe University. Turkey.

Betz, N. E. & Weiss, D. J. (1974). *Simulation studies of two stage ability testing. Research report*. Research Report 74-4. Minneapolis: University of Minnesota. Psychometric Methods Program. Department of Psychology.

Binet, A. & Simon, T. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191-244.

Birnbaum, A. (1958). *On the estimation of mental ability* (Series Report No. 15. Project No. 7755-23). Randolph Air Force Base TX: USAF School of Aviation Medicine.

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

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

Bock, R. D. & Mislevy, R. J. (1981). *Data quality analysis of the Armed Services Vocational Aptitude Battery*. Chicago: National Opinion Research Center.

Bock, R. D. & Mislevy. R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement, 6*, 431-444.

Çalışkan, M. (2000). *The fit of one-. two- and three-parameter models of item response theory to the ministry of national education-educational research and development directorate's science achievement*. Unpublished Master Thesis. Middle East Technical University. Department of Educational Sciences.

Chang, S, W. & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

Cikrikci-Demirtaslı, N. (1999). Psikometride yeni ufuklar: bilgisayar ortamında bireye uyarlanmış test. *Türk Psikoloji Bülteni, 5*(13), 31-36.

De Ayala, R. J. (1992). Nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327-343.

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

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement, 41,* 261–270.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Economides, A. A. & Roupas, C. (2007). Evaluation of Computer Adaptive Testing Systems. *International Journal of Web-Based Learning and Teaching Technologies, 2*(1), 70-87.

Eggen. T.J.H.M. & Straetmans. G.J.J.M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement. 60*(5), 713-734

Embretson, S. E. (1996). The New Rules of Measurement. *Psychological Assessment, 8*(4), 341-349.

Embretson, S. E. & Reise., S. P. (2000). I*tem response theory for psychologists*. Mahwah. NJ. Erlbaum.

Engelhard, E. (1986). A simulation study of computerized adaptive testing with a misspecified measurement model. Paper presented at the annual conference of the American Statistical Association.

Ertkin, E. (1993). *Geleneksel .lçme kuramına alternatif iki yöntemin tanıtılması ve personel seçimine yönelik uygulama çalışması*. Unpublished doctoral dissertation, Istanbul University, Turkey.

Ferrando, P. J. & Lorenzo-Seva, U. (2001). Observed scores: checking the appropriateness of item response theory models by predicting the distribution of observed scores: the program eo-fit. *Educational and Psychological Measurement, 61,* 895-902.

Gelbal, S. (1994). *p madde gucluk indeksi ile rasch modelinin b parametresi ve bunlara dayali yetenek olculeri uzerine bir karsilastirma*. Unpublished doctoral dissertation. Hacettepe University., Turkey.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B. Klapp, B. F. & Rose. M. (2005). D*evelopment of a computer-adaptive test for depression (D-CAT). Quality of Life Research, 14*, 2277-2291.

GMAT Official Web Site. Retrived from http://www.mba.com/mba/thegmat. (Last visited on 25/12/2010)

GRE Official Web Site. Retrived from http://www.ets.org/gre/. (Last visited on 25/12/2010)

Gushta, M. M. (2003). *Standard-setting issues in computerized-adaptive testing.* Paper Prepared for Presentation at the Annual Conference of the Canadian Society for Studies in Education. Halifax. Nova Scotia. May 30th, 2003.

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. & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.). *Applications of item response theory* (pp. 71-94). Vancouver. British Columbia: Education Research Institute of British Columbia.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M.(1991). Computerized adaptive testing: Theory. applications standards. In R. K. HAmbleton & J. N. Zaal (Eds.). *Advances in educational and psychological testing: Theory and applications.* Boston: Kluwer Academic Publishers.

Hambleton, R. K. & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

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

Han, K. T. (2007). Wingen: windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement, 31*(5), 457–459.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133–46.

Harwell, M., Stone, C. A., Hsu, T. C. & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-126.

Iseri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures*. Unpublished doctoral dissertation. Middle East Technical University, Turkey.

ITEMAN. (2010). Iteman: Item Analysis Software. [Computer software]. Chicago: Scientific Software International.

Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireysellestirilmis) test uygulamasi ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması*. Unpublished doctoral dissertation, Hacettepe University. Turkey.

Koklu, N. (1990). Klasik test teorisinie gore gelistirilen tailored test ile grup testi arasında bir karşılaştırma. Unpublished doctoral dissertation. Hacettepe University, Turkey.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: methods and practices*. New York: Springer-Verlag.

Legg, S. M. & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*(2), 23-27.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Liang, T., Han, K. T. & Hambleton, R. K. (2009). Residplots-2 computer software for irt graphical residual analyses. *Applied Psychological Measurement, 33*(5), 411-412.

Lord, F. M. (1952). *A theory of test scores*. Psychometric monograph. No: 7.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood of item bias in a test of reading comprehension. *Applied Psychological Measurement, 18,* 57-75.

Lord, F. M. (1980). *An application of item response theory to practical testing problems*. Hillsdale. NJ: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157-162.

Lord, F. N. & Novick. M. R. (1968). *Statistical theories of mental test scores*. Reading. MA: Addison-Wesley.

Ludlow, L. H. (1986). Graphical Analysis of Item Response. *Applied Psychological Measurement, 10*, 217-229.

Lunz, M. E. & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31*(2), 251-263.

MALT Official Web Site. Retrieved from http://www.hoddertests.co.uk/tfsearch/ks3/numeracy/MaLTnew.htm. (Last visited on 25/12/2010)

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military design. In D. J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

McKinley, R. & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 19*, 49-57.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin 1993, 114*(3), 449-458.

Miller, V. D. (2003). *Assessment of student achievement: a comparative study of student achievement using paper and pencil assessment and computerized adaptive testing*. Unpublished dissertation. Wayne State University. Graduate School.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287-304.

NWEA Official Web Site. Retrieved from http://www.nwea.org. (Last visited on 25/12/2010)

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton NJ: Educational Testing Service.

Papanastasiou, E. (2003). *Computer-adaptive testing in science education*. Paper presented at 6th International Conference on Computer Based Learning in Science. 965-971.

Raîche, G. & Blais, J.(2002). *Practical considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori. adaptive correction for bias. and adaptive integration interval*. Paper presented at the Biennal International Objective Measurement Workshop New Orleans.

Rasch, G. (1960). *Probabilistic models for some intellignece and attainment tests*. Copanhagen, Denmark. Danish Institute for Educational Research.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143–151.

Riley, B.B., Conrad, K.J., Bezruczko, N., Dennis, M. (2007). Relative Precision, Efficiency and Construct Validity of Different Starting and Stopping Rules for a Computerized Adaptive Test: The GAIN Substance Problem Scale. *Journal of Applied Measurement*, *8*(1).

Rudman, H. C. (1987). The future of testing is now. Educational Measurement: *Issues and Practice, 4*, 5-11.

Rudner, L. M. (1998). An on-line. Interactive Computer Adaptive Testing Mini Tutorial [Online]. Retrieved from http://edres.org/scripts/cat/catdemo.htm. (Last visited on 25/12/2010)

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometric Monograph*, No.17.

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

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). *Psychological Assessment, 17,* 28-43.

SMI Official Web Site. (2010) Retrieved from http://www.scholastic.com/readeveryday. (Last visited on 01/11/2010)

Stone, C. A. (2004). IRTFIT−RESAMPLE: A Computer Program for Assessing Goodness of Fit of Item Response Theory Models Based on Posterior Expectations. *Applied Psychological Measurement, 28*(2), 143–144.

Stroud, A. H. & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs. NJ: Prentice-Hall.

Student Selection and Placement Center. (2010). Retrieved from http://www.osym.gov.tr. (Last visited on 01/11/2010)

TOEFL Official Web Site. (2010). Retrieved from http://www.ets.org/toefl.(Last visited on 01/11/2010)

Trabin, T. E. & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.). *New horizons in testing*. New York: Academic Press.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]*. Lisse: Swets & Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267–298.

van der Linden, W.J. & Hambleton, R.K. (Eds.) (1996). *Handbook of modern item response theory*. New York: Springer.

van der Linden, W. J. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika. 73*(3), 365-384.

Wainer, H., Dorans, N., Flaughter,. R., Green, B., Mislevy, R., Steinberg, L. & Thissen, D. (1990) *Computerized adaptive testing: A primer*. Hillsdale. NJ: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (1990). Item response theory. item calibration. and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg & D. Thissen. *Computerized adaptive testing: A primer* (pp.65–101). Hillsdale NJ: Erlbaum.

Wang, T. (1997. March). *Essentially unbiased eap estimates in computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association. Chicago IL.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135.

Wang, X., Bo-Pan, W. & Harris, V. (1999). *Computerized Adaptive Testing Simulations Using Real Test Taker Responses*. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109-135.

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.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weiss, David J. (1983). Latent trait theory and adaptive testing. In David J. Weiss (ed.). *New horizons in testing* (pp. 5-7). New York: Academic Press.

Weiss, D. J. (2010). CAT Central: *A global resource for Computerized Adaptive Testing Research and Applications* [Online]. http://www.psych.umn.edu/psylabs/CATCentral. Last visited on 25/12/2010.

Weiss, D. J. & Betz. N. E. (1973). *Ability measurement: conventional or adaptive?*. Research report. 73-1.

Whitely, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.

Wilson, D. T., Wood, R. & Gibbons, R. (1991). *Testfact test scoring, item statistics, and item factor analysis*. [Computer software]. Chicago: Scientific Software International.

Wise, S. L. & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18(2),* 163-183.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D. & Mead, R. J. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago. Statistical Laboratory, Department of Education.

Yasar, M. (1999). *Bireysellestirilmis testler uzerine bir calisma*. Unpublished doctoral dissertation, Hacettepe University. Turkey.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yenal, E. (1995). *Differential item functioning analysis of the quantitative ability section of the first stage of the university entrance examination in turkey*. Unpublished doctoral dissertation. Middle East Technical University, Turkey.

Yi, Q., Wang, T., & Ban, J. C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38,* 267-292.

Yildirim, H. H., Çömlekoğlu, G., Berberoğlu, G. (2003). Milli eğitim bakanlığı özel okullar sınavı verilerinin madde tepki kuramı modellerine uyumu. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 23,* 159-168.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.

APPENDIX A

IRT PARAMETERS OF LIVE CAT ITEM BANK

Table A.1 Item Parameters Descriptive for Live CAT

	A	b	c
Mean	0.990	0.889	0.017
Median	0.883	0.817	0.018
Mode	0.630	0.000	0.018
Std. Deviation	0.500	0.730	0.019
Variance	0.250	0.533	0.000
Range	3.440	4.663	0.155
Minimum	0.259	-1.603	0.000
Maximum	3.699	3.060	0.155

Table A.2 Item Parameters for Live CAT

#	Year	Item # in Booklet	a	b	c
1	2000	46	0.948	0.256	0.018
2	2000	47	0.925	-0.607	0.018
3	2000	48	0.925	0.148	0.018
4	2000	49	1.500	0.151	0.018
5	2000	51	0.614	1.230	0.018
6	2000	52	0.905	0.959	0.018
7	2000	53	0.674	-0.361	0.018
8	2000	55	0.831	0.396	0.018
9	2000	58	0.858	0.941	0.018
10	2000	59	0.623	2.535	0.018
11	2000	60	0.865	0.631	0.018
12	2000	61	1.003	0.621	0.018
13	2000	62	0.633	0.425	0.018
14	2000	63	0.972	0.000	0.018
15	2000	64	0.935	0.942	0.018
16	2000	65	0.890	0.876	0.018
17	2000	66	0.411	1.944	0.018

Table A. 2 continued

18	2000	67	0.716	-0.259	0.018
19	2000	68	0.587	1.271	0.018
20	2000	69	1.230	0.995	0.018
21	2000	70	0.694	0.309	0.018
22	2000	71	0.623	0.577	0.018
23	2000	72	1.293	-0.255	0.018
24	2000	73	0.341	1.448	0.018
25	2000	74	1.017	0.991	0.018
26	2000	75	0.630	1.325	0.018
27	2000	76	0.951	0.405	0.018
28	2000	77	1.259	0.986	0.018
29	2000	78	0.642	0.866	0.018
30	2000	79	0.541	1.042	0.018
31	2000	80	0.591	1.517	0.018
32	2000	82	0.453	0.000	0.018
33	2000	83	0.461	0.667	0.018
34	2000	84	0.774	1.318	0.018
35	2000	85	0.607	2.707	0.018
36	2000	86	0.630	2.404	0.018
37	2000	87	0.403	1.639	0.018
38	2000	88	0.541	2.692	0.018
39	2000	89	0.687	0.540	0.018
40	2000	90	0.752	-0.596	0.018
41	2001	46	0.964	1.375	0.018
42	2001	47	0.478	0.175	0.018
43	2001	48	0.602	0.097	0.018
44	2001	49	1.017	0.035	0.018
45	2001	50	0.531	1.180	0.018
46	2001	51	0.754	0.125	0.018
47	2001	52	0.679	0.314	0.018
48	2001	53	0.738	0.603	0.018
49	2001	54	0.967	1.063	0.018
50	2001	55	0.876	-0.384	0.018
51	2001	56	0.426	-0.064	0.018
52	2001	57	0.406	2.145	0.018
53	2001	58	1.068	0.069	0.018
54	2001	59	0.818	0.040	0.018
55	2001	60	1.124	0.480	0.018
56	2001	61	1.816	0.700	0.018
57	2001	62	0.527	1.806	0.018
58	2001	63	0.855	0.511	0.018
59	2001	64	0.493	3.033	0.018

Table A. 2 continued

60	2001	65	0.862	1.033	0.018
61	2001	66	1.494	0.850	0.018
62	2001	67	0.905	0.782	0.018
63	2001	68	0.496	1.380	0.018
64	2001	69	0.807	0.976	0.018
65	2001	70	0.840	-1.309	0.018
66	2001	71	1.180	0.884	0.018
67	2001	72	0.562	0.256	0.018
68	2001	73	1.113	1.282	0.018
69	2001	74	1.068	0.138	0.018
70	2001	75	0.964	0.927	0.018
71	2001	76	0.630	1.641	0.018
72	2001	78	0.659	0.046	0.018
73	2001	80	0.510	0.731	0.018
74	2001	81	0.555	2.642	0.018
75	2001	82	0.727	1.763	0.018
76	2001	84	0.735	1.751	0.018
77	2001	85	0.259	2.205	0.018
78	2001	87	0.484	2.813	0.018
79	2001	90	0.559	2.038	0.018
80	2002	46	0.655	1.231	0.018
81	2002	47	0.351	2.883	0.018
82	2002	48	0.424	1.650	0.018
83	2002	49	0.504	1.165	0.018
84	2002	50	0.402	0.541	0.018
85	2002	51	0.669	-0.409	0.018
86	2002	52	0.727	0.843	0.018
87	2002	53	0.450	1.006	0.018
88	2002	54	0.813	0.568	0.018
89	2002	55	0.555	0.850	0.018
90	2002	56	0.514	1.546	0.018
91	2002	57	0.573	1.357	0.018
92	2002	58	1.000	-0.178	0.018
93	2002	59	0.790	0.578	0.018
94	2002	60	1.097	0.710	0.018
95	2002	61	0.645	1.021	0.018
96	2002	62	0.453	1.484	0.018
97	2002	63	0.959	0.886	0.018
98	2002	65	0.614	1.542	0.018
99	2002	66	1.053	0.887	0.018
100	2002	67	0.803	0.201	0.018
101	2002	68	1.053	0.494	0.018

Table A. 2 continued

102	2002	69	0.475	1.359	0.018
103	2002	70	0.735	0.838	0.018
104	2002	71	1.177	0.885	0.018
105	2002	72	0.956	0.292	0.018
106	2002	73	1.632	0.250	0.018
107	2002	74	0.683	1.492	0.018
108	2002	75	0.980	0.964	0.018
109	2002	76	0.790	2.380	0.018
110	2002	77	1.094	0.672	0.018
111	2002	78	0.935	0.897	0.018
112	2002	80	0.344	3.060	0.018
113	2002	81	0.429	1.331	0.018
114	2002	82	0.471	1.893	0.018
115	2002	83	0.685	1.620	0.018
116	2002	84	0.362	2.271	0.018
117	2002	86	0.762	1.710	0.018
118	2002	87	0.618	0.384	0.018
119	2002	90	0.627	0.237	0.018
120	2004	46	0.710	-0.393	0.018
121	2004	47	0.770	0.588	0.018
122	2004	48	0.959	0.000	0.018
123	2004	49	1.211	0.098	0.018
124	2004	50	0.521	2.543	0.018
125	2004	51	0.750	0.422	0.018
126	2004	52	0.637	0.768	0.018
127	2004	54	1.032	0.854	0.018
128	2004	55	0.740	0.296	0.018
129	2004	56	0.803	1.462	0.018
130	2004	57	0.593	0.098	0.018
131	2004	59	0.975	0.036	0.018
132	2004	60	1.114	1.025	0.018
133	2004	61	0.567	-0.673	0.018
134	2004	62	0.964	1.162	0.018
135	2004	63	0.703	1.464	0.018
136	2004	64	0.820	0.441	0.018
137	2004	65	0.601	1.931	0.018
138	2004	66	1.020	0.319	0.018
139	2004	67	1.600	0.585	0.018
140	2004	68	1.324	0.483	0.018
141	2004	69	0.833	1.836	0.018
142	2004	70	0.733	1.365	0.018
143	2004	71	1.144	0.268	0.018

Table A. 2 continued

144	2004	72	0.660	2.044	0.018
145	2004	73	0.831	1.622	0.018
146	2004	74	0.617	-1.603	0.018
147	2004	75	0.917	0.337	0.018
148	2004	76	0.977	0.362	0.018
149	2004	77	0.824	0.277	0.018
150	2004	78	0.723	0.345	0.018
151	2004	79	0.735	0.516	0.018
152	2004	81	0.750	0.733	0.018
153	2004	82	0.688	0.222	0.018
154	2004	84	0.375	2.833	0.018
155	2004	85	0.648	0.911	0.018
156	2004	86	0.295	0.623	0.018
157	2004	88	0.418	1.359	0.018
158	2004	89	0.549	1.825	0.018
159	2004	90	0.483	1.935	0.018
160	2005	48	3.699	0.395	0.033
161	2005	50	1.564	0.779	0.001
162	2005	51	1.309	1.244	0.005
163	2005	55	0.776	1.610	0.000
164	2005	56	1.322	0.780	0.000
165	2005	57	2.129	0.471	0.130
166	2005	58	1.511	0.633	0.000
167	2005	59	2.620	0.419	0.012
168	2005	60	1.066	0.661	0.000
169	2005	62	1.410	0.882	0.017
170	2005	63	2.562	0.334	0.126
171	2005	64	0.937	1.045	0.000
172	2005	66	2.544	0.428	0.000
173	2005	67	1.300	0.822	0.000
174	2005	68	1.328	1.067	0.000
175	2005	70	1.009	1.168	0.000
176	2005	71	1.047	1.423	0.000
177	2005	73	0.965	1.441	0.001
178	2005	75	1.795	0.626	0.001
179	2005	76	2.046	0.548	0.001
180	2005	77	1.422	0.670	0.000
181	2005	78	1.685	0.802	0.000
182	2005	79	1.943	0.593	0.008
183	2005	80	1.485	0.695	0.046
184	2005	81	1.484	0.790	0.000
185	2005	82	0.816	1.675	0.000

Table A. 2 continued

186	2005	83	1.685	1.037	0.063
187	2005	84	1.926	0.752	0.002
188	2005	85	0.890	1.387	0.000
189	2005	86	0.853	1.559	0.007
190	2005	87	1.658	0.594	0.000
191	2005	88	0.813	1.500	0.000
192	2005	89	1.870	0.759	0.011
193	2005	90	1.420	0.662	0.001
194	2006	1	1.693	0.379	0.138
195	2006	2	1.706	0.521	0.036
196	2006	4	1.147	1.112	0.000
197	2006	5	1.468	0.836	0.025
198	2006	6	1.519	0.712	0.025
199	2006	8	1.185	1.012	0.036
200	2006	9	1.581	0.656	0.020
201	2006	10	0.849	1.838	0.008
202	2006	11	1.026	1.340	0.000
203	2006	12	1.036	0.648	0.007
204	2006	13	1.331	0.419	0.002
205	2006	14	1.549	0.860	0.015
206	2006	15	1.157	1.170	0.015
207	2006	16	1.585	0.906	0.002
208	2006	18	1.431	0.928	0.000
209	2006	19	0.944	1.254	0.000
210	2006	21	0.825	1.535	0.000
211	2006	22	1.158	1.173	0.000
212	2006	23	1.046	0.611	0.005
213	2006	24	0.981	1.060	0.000
214	2006	26	1.195	0.725	0.042
215	2006	27	1.262	1.262	0.016
216	2007	1	0.782	1.675	0.000
217	2007	2	0.835	0.854	0.000
218	2007	3	1.268	0.812	0.000
219	2007	4	0.658	1.517	0.000
220	2007	5	0.969	0.537	0.001
221	2007	6	2.293	0.371	0.045
222	2007	7	1.449	0.763	0.026
223	2007	8	1.943	0.480	0.028
224	2007	9	1.236	0.446	0.009
225	2007	10	1.727	0.351	0.012
226	2007	11	1.265	0.328	0.098
227	2007	12	1.265	0.342	0.002

Table A. 2 continued

228	2007	13	0.924	1.252	0.000
229	2007	14	1.394	0.688	0.000
230	2007	17	1.585	0.475	0.000
231	2007	19	2.251	0.688	0.000
232	2007	20	0.981	1.625	0.002
233	2007	21	2.320	0.591	0.000
234	2007	22	1.853	0.619	0.000
235	2007	23	1.373	0.930	0.155
236	2007	24	1.136	0.875	0.000
237	2007	25	0.892	1.112	0.027
238	2007	26	0.862	0.924	0.023
239	2007	28	1.528	0.390	0.048
240	2007	29	0.475	-0.298	0.003
241	2007	30	0.936	0.597	0.000
242	2007	46	2.930	0.413	0.036

APPENDIX B

SCORES OF EXAMINEES FROM P&P AND CAT

Table B.1 Live Testing Examinees' Scores

		P&P SSE			CAT SSE		
#	School	# of Given	Theta	SE	# of Given	Theta	SE
1	Anatolian	30	0.493	0.1658	4	0.53	0.2858
2	Anatolian	30	-0.6311	0.4361	4	0.5	0.269
3	State	30	-0.2054	0.2715	7	0.2	0.225
4	Anatolian	30	0.4775	0.17	14	0.04	0.2654
5	Anatolian	30	-0.9894	0.642	17	-0.27	0.297
6	State	30	2.4861	0.521	13	1.34	0.2878
7	Anatolian	30	2.5743	0.5889	15	1.39	0.2833
8	Anatolian	30	1.9291	0.3573	7	0.99	0.29
9	Anatolian	30	2.4784	0.5184	18	1.68	0.2913
10	State	30	2.9671	0.7685	19	1.78	0.2985
11	Anatolian	30	2.2186	0.4356	30	2.82	0.3818
12	Anatolian	30	3.0969	0.7807	27	2.17	0.2995
13	Anatolian	30	3.1335	0.7999	17	1.62	0.2984
14	State	30	3.5484	0.9931	30	2.66	0.3731
15	State	30	1.1421	0.2145	8	0.91	0.2665
16	Anatolian	30	2.0247	0.3817	9	1.03	0.2709
17	Anatolian	30	2.5172	0.5319	7	0.75	0.2475
18	Anatolian	30	2.0887	0.3989	9	0.88	0.2683
19	Anatolian	30	0.1648	0.1949	4	0.25	0.2961
20	Anatolian	30	1.2064	0.2148	7	0.84	0.2963
21	State	30	2.9671	0.7685	8	0.95	0.2766
22	Anatolian	30	3.0969	0.7807	8	1.03	0.2752
23	Anatolian	30	2.5446	0.5416	18	1.46	0.2756

Table B.1 continued

24	State	30	2.4651	0.5471	27	1.97	0.2916
25	Anatolian	30	1.7109	0.3256	8	0.44	0.2281
26	Anatolian	30	0.1538	0.1964	13	0.04	0.2649
27	Anatolian	30	2.0081	0.3774	7	0.82	0.2791
28	Private	30	2.5358	0.5385	4	0.26	0.2872
29	Anatolian	30	2.5554	0.5455	14	1.29	0.2798
30	Anatolian	30	0.537	0.1685	4	0.26	0.2915
31	Anatolian	30	2.3273	0.4987	6	0.37	0.2397
32	Anatolian	30	2.5144	0.5309	20	1.52	0.299
33	Anatolian	30	0.358	0.176	5	0.16	0.2878

APPENDIX C

IRT PARAMETERS OF SIMULATIONS

Table C.1 IRT Parameters for Post-Hoc Simulation for 2005 State Schools

#	a	b	c	#	a	b	c
1	3.609	0.626	0.033		25	1.484	1.702
2	1.469	2.489	0.002	26	1.401	1.904	0.006
3	3.66	0.612	0.007	27	2.029	1.101	0
4	3.974	0.431	0.034	28	1.527	1.874	0
5	1.953	1.189	0.009	29	1.758	1.416	0
6	2.138	1.686	0.007	30	2.402	0.894	0.015
7	3.639	0.932	0	31	2.669	0.828	0.005
8	3.348	0.938	0.008	32	1.873	1.018	0
9	1.771	0.977	0.03	33	2.442	1.035	0
10	1.084	2.194	0.001	34	2.445	0.735	0.004
11	1.897	1.151	0.022	35	1.766	0.811	0.003
12	2.66	0.683	0.027	36	1.88	1.147	0
13	1.833	0.871	0	37	1.097	2.321	0
14	3.265	0.638	0.022	38	2.014	1.447	0.047
15	1.564	0.965	0.001	39	2.137	1.107	0
16	2.733	1.49	0.015	40	1.241	2.033	0
17	1.359	1.353	0	41	1.267	2.133	0
18	3.4	0.479	0.082	42	1.953	0.966	0.019
19	1.431	1.243	0.001	43	1.165	1.997	0
20	1.53	0.841	0.018	44	1.649	1.268	0
21	3.342	0.596	0.047	45	1.756	0.946	0
22	1.49	1.05	0.001				
23	1.794	1.465	0				
24	2.822	0.926	0				

Table C.2 IRT Parameters for Post-Hoc Simulation for 2005 Anatolian Schools

#	a	b	c	#	a	b	C
1	1.338	-2.817	0.001	26	1.652	-0.023	0
2	2.442	0.545	0.019	27	1.356	-0.756	0.001
3	1.72	-1.698	0.001	28	1.141	0.295	0
4	1.797	-2.216	0.001	29	1.324	-0.234	0
5	1.276	-0.78	0.001	30	1.867	-1.01	0.001
6	1.641	-0.14	0	31	1.961	-1.25	0.002
7	2.455	-1.228	0	32	1.665	-1.027	0.001
8	1.512	-1.163	0.002	33	1.766	-0.926	0
9	1.197	-1.025	0.001	34	2.151	-1.074	0
10	1.874	0.458	0.023	35	2.077	-0.759	0
11	1.126	-0.976	0.001	36	1.353	-0.67	0.001
12	2.097	-1.492	0.001	37	1.175	0.649	0.005
13	1.352	-1.023	0.001	38	2.413	-0.45	0.001
14	1.714	-1.843	0	39	1.781	-0.822	0
15	1.142	-0.9	0.001	40	1.47	0.079	0
16	2.234	-0.322	0.007	41	1.479	0.25	0.011
17	2.077	-0.455	0.029	42	1.887	-0.75	0
18	1.647	-2.062	0.003	43	1.052	0.371	0
19	0.911	-0.586	0.001	44	2.024	-0.952	0.023
20	1.324	-1.367	0.001	45	1.392	-1.047	0.001
21	1.815	-1.703	0.05				
22	1.16	-0.921	0.001				
23	1.277	-0.438	0.001				
24	3.04	-0.85	0				
25	0.998	-0.109	0				

Table C.3 IRT Parameters for Post-Hoc Simulation for 2005 Private Schools

#	a	b	c	#	a	b	c
1	1.237	-0.825	0.057	26	1.353	0.556	0.014
2	1.652	0.973	0.006	27	1.303	0.039	0
3	1.661	-0.584	0.023	28	0.898	0.812	0.002
4	1.829	-0.869	0.032	29	1.253	0.2	0
5	1.281	-0.086	0.006	30	1.245	-0.309	0.001
6	1.253	0.536	0.002	31	1.764	-0.42	0.013
7	1.96	-0.368	0.025	32	1.233	-0.354	0
8	1.453	-0.145	0.016	33	1.49	-0.197	0.011
9	0.97	-0.3	0.001	34	1.668	-0.365	0.052
10	1.363	0.795	0.025	35	1.403	-0.202	0.035
11	1.109	-0.052	0.022	36	1.204	0.034	0
12	1.568	-0.563	0.064	37	0.992	0.808	0.004
13	1.077	-0.369	0	38	1.73	0.125	0
14	1.497	-0.718	0	39	1.515	-0.042	0.001
15	0.954	-0.043	0	40	1.264	0.665	0
16	1.901	0.291	0	41	1.267	0.703	0.011
17	1.451	0.113	0	42	1.444	-0.1	0.013
18	1.387	-0.896	0.092	43	0.879	0.86	0
19	0.991	0.191	0	44	1.653	-0.095	0.013
20	1.081	-0.516	0.001	45	1.16	-0.218	0.005
21	1.865	-0.574	0.046				
22	1.015	0.038	0.032				
23	1.246	0.275	0				
24	2.345	-0.133	0.012				
25	1.036	0.479	0				

Table C.4 IRT Parameters for Post-Hoc Simulation for 2006 State Schools

#	a	b	c
1	2.567	0.488	0.050
2	2.590	0.769	0.019
3	1.314	1.201	0.000
4	1.399	1.528	0.000
5	1.948	1.357	0.008
6	1.461	1.128	0.001
7	1.032	3.712	0.024
8	0.923	2.035	0.001
9	1.988	1.014	0.001
10	1.044	2.526	0.000
11	1.268	2.077	0.000
12	1.454	1.066	0.020
13	2.364	0.606	0.004
14	2.896	1.051	0.022
15	1.332	1.860	0.010
16	3.540	1.364	0.016
17	2.208	1.113	0.001
18	2.703	1.136	0.009
19	1.437	1.735	0.000
20	1.393	1.761	0.000
21	0.932	2.525	0.000
22	1.940	1.530	0.001
23	1.711	0.916	0.009
24	1.475	1.548	0.000
25	0.790	2.954	0.001
26	2.024	1.093	0.044
27	1.373	2.049	0.004
28	0.466	1.447	0.002
29	0.859	3.612	0.000
30	0.603	2.168	0.003

Table C.5 IRT Parameters for Post-Hoc Simulation for 2006 Anatolian Schools

#	a	b	c
1	1.097	-1.934	0.001
2	2.523	-0.477	0.014
3	1.338	0.063	0.001
4	1.698	0.189	0.000
5	2.347	-0.265	0.006
6	1.843	-0.369	0.000
7	2.753	0.553	0.011
8	2.204	0.202	0.019
9	2.180	-0.286	0.000
10	2.854	0.774	0.000
11	2.244	0.345	0.000
12	1.930	-0.180	0.056
13	1.678	-0.779	0.001
14	2.097	-0.291	0.002
15	3.556	0.287	0.005
16	3.431	-0.032	0.008
17	2.684	-0.194	0.022
18	2.484	-0.088	0.000
19	2.109	0.438	0.000
20	2.274	0.332	0.000
21	2.215	0.391	0.004
22	2.105	0.101	0.000
23	2.133	-0.399	0.001
24	2.468	0.093	0.000
25	2.388	0.625	0.001
26	2.261	-0.291	0.002
27	2.723	0.362	0.046
28	0.653	-0.758	0.001
29	0.702	4.181	0.000
30	0.854	0.720	0.001

Table C.6 IRT Parameters for Post-Hoc Simulation for 2006 Private Schools

#	a	b	c
1	1.102	-0.725	0.001
2	1.517	-0.150	0.001
3	0.965	0.525	0.000
4	1.199	0.563	0.000
5	1.955	0.274	0.031
6	1.459	0.173	0.001
7	2.622	0.918	0.015
8	1.307	0.621	0.000
9	1.620	0.188	0.000
10	1.890	1.058	0.009
11	1.626	0.664	0.000
12	1.374	0.229	0.082
13	1.287	-0.148	0.000
14	1.845	0.303	0.009
15	1.836	0.702	0.011
16	2.612	0.490	0.003
17	2.271	0.287	0.044
18	1.857	0.383	0.004
19	1.283	0.839	0.000
20	1.569	0.697	0.004
21	1.987	0.788	0.020
22	1.612	0.597	0.000
23	1.556	0.178	0.000
24	1.755	0.545	0.010
25	1.827	0.974	0.011
26	1.366	0.231	0.014
27	1.693	0.757	0.002
28	0.975	0.649	0.226
29	0.608	3.500	0.000
30	0.509	1.115	0.002

APPENDIX D

HISTROGRAMS OF IRT PARAMETERS

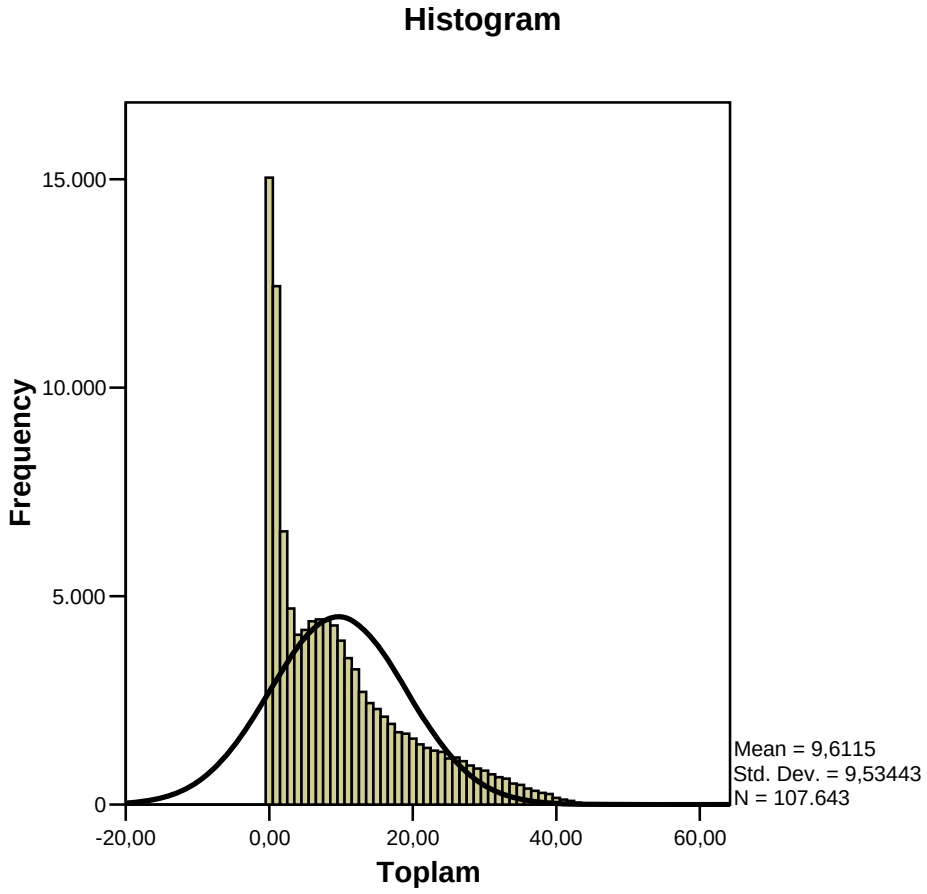

Figure D.1 Ability Distributions of 2005 State High Schools

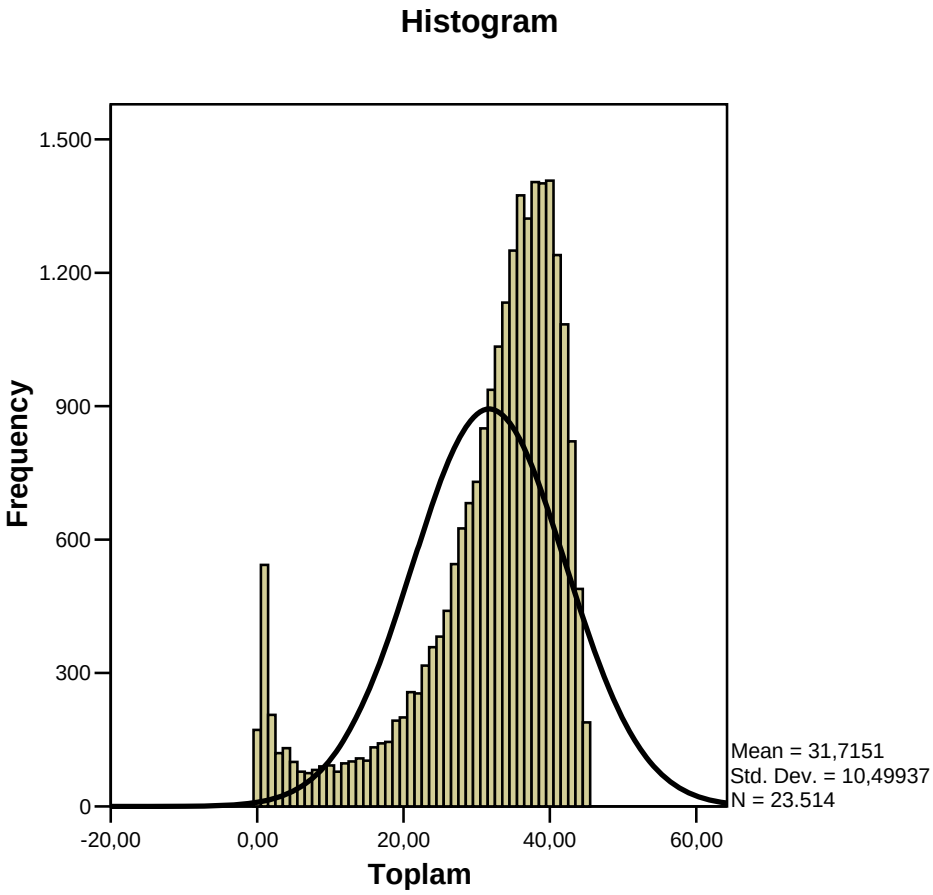

Figure D.2 Ability Distributions of 2005 Anatolian High Schools

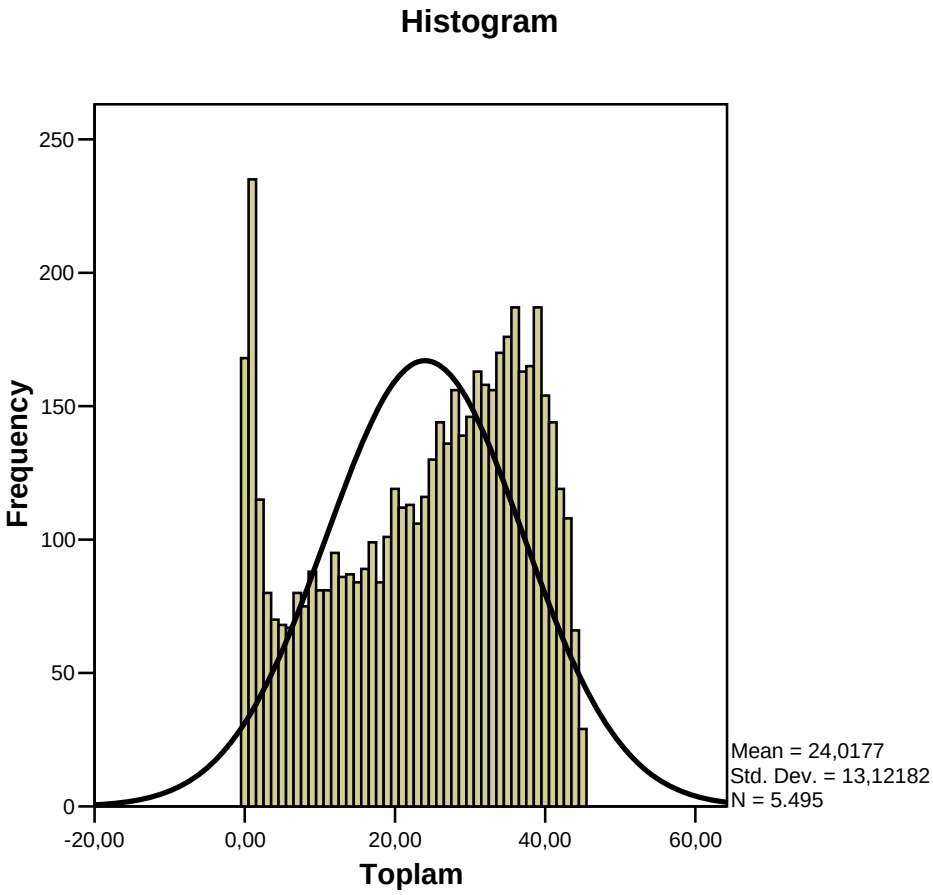

Figure D.3 Ability Distributions of 2005 Private High Schools

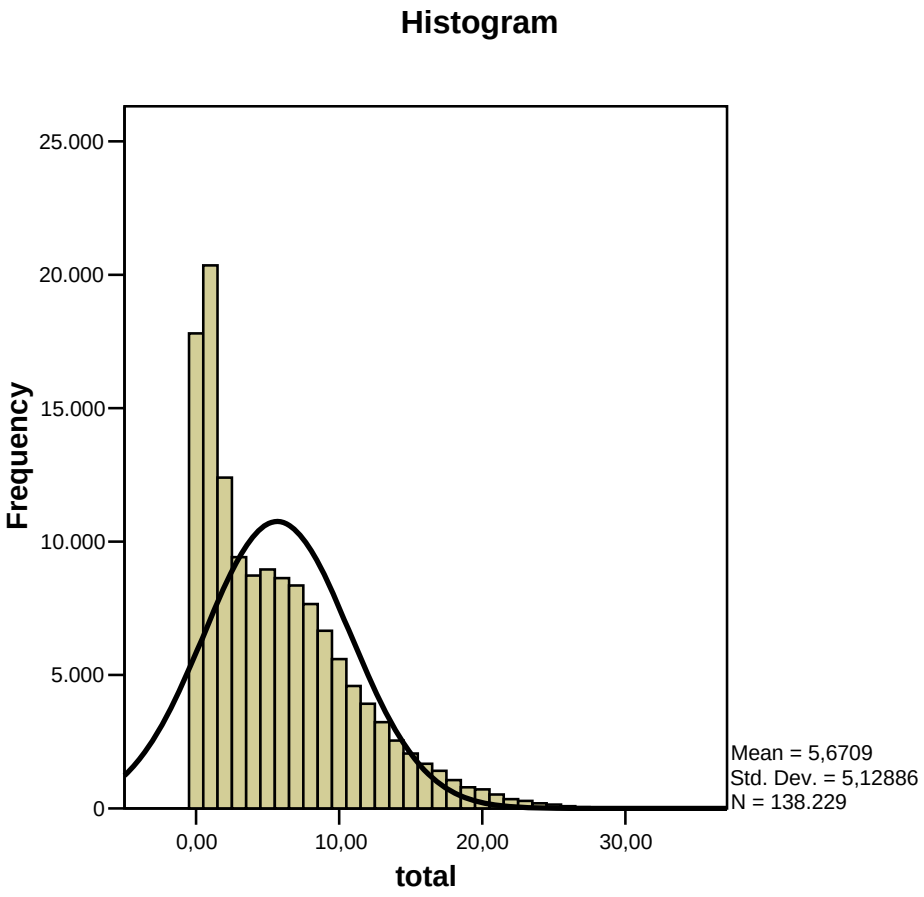

Figure D.4 Ability Distributions of 2006 State High Schools

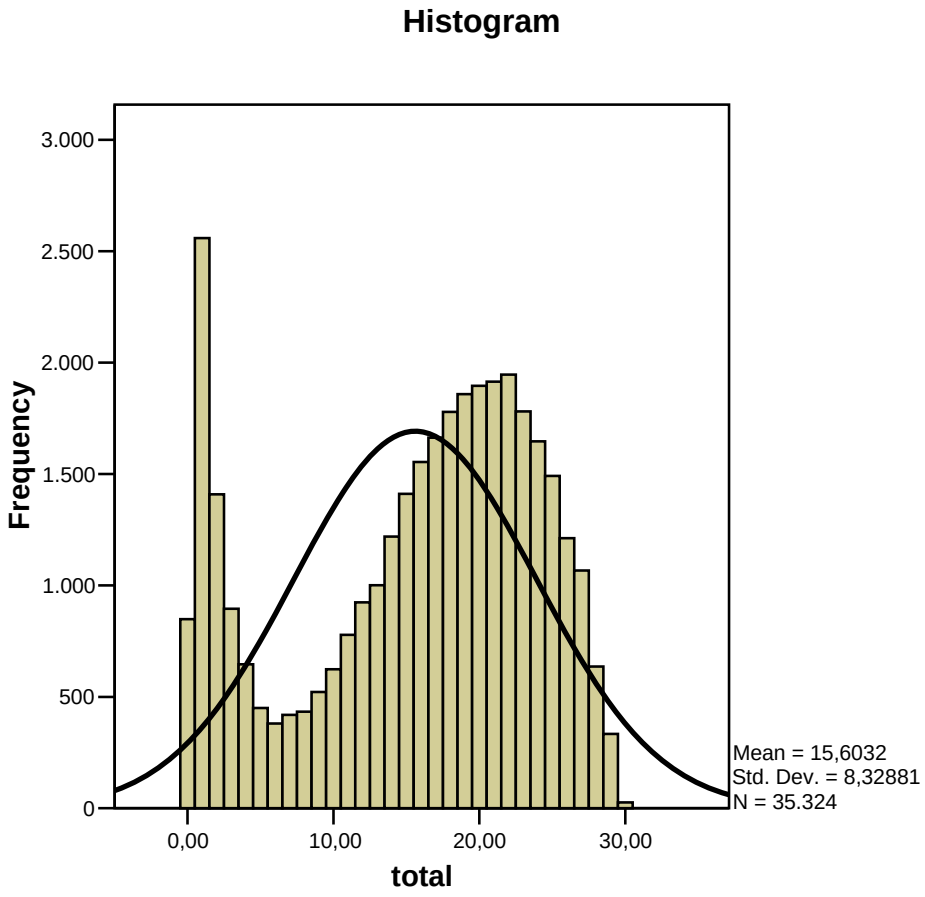

Figure D.5 Ability Distributions of 2006 Anatolian High Schools

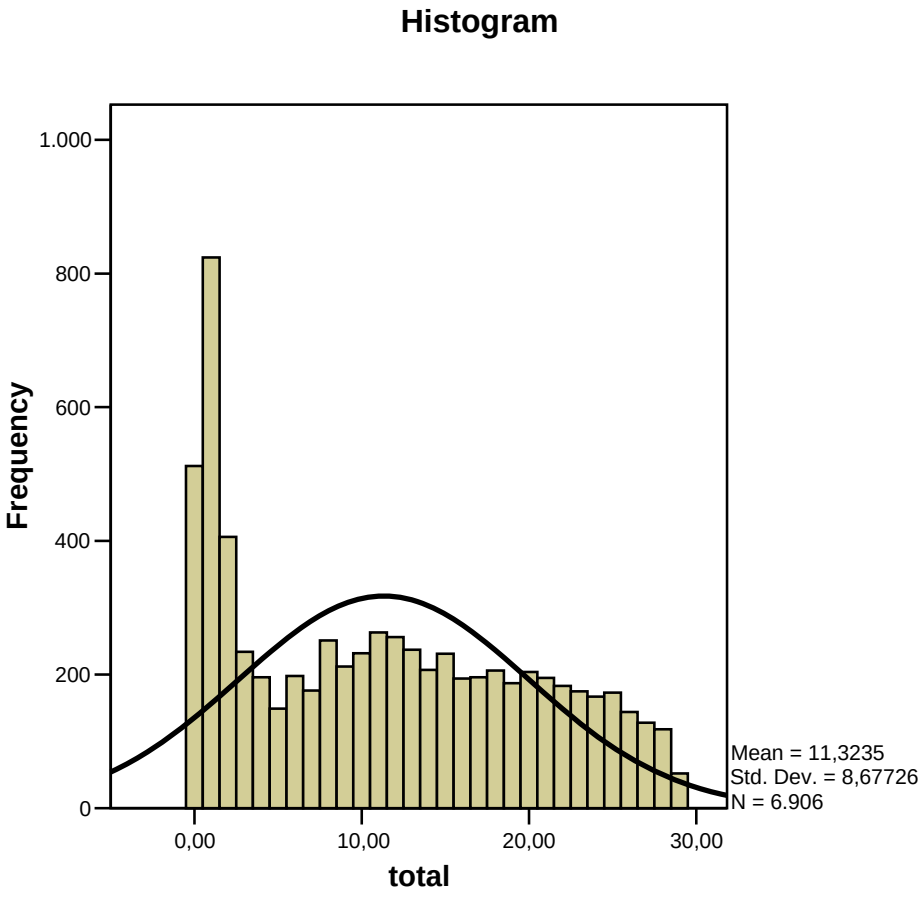

Figure D.6 Ability Distributions of 2006 Private High Schools

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Kalender, Ilker
Nationality: Turkish (TC)
Date and Place of Birth: April, 19 1978, Istanbul, Turkey
Marital Status: Married
Phone: +90 506 534 6401
Email: kalenderi@bilkent.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Secondary Sci. and Math.Edu.	2004
BS	METU Physics	2002
High School	Sakip Sabanci İstanbul	1995

WORK EXPERIENCE

Year	Place	Enrollment
2008- Present	Bilkent University	Instructor
2006-2008	Midde East Technical University	Research Assistant

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

- Kalender, I. & Berberoglu, G. (2009). An assessment of factors related to science achievement of turkish students. *International Journal of Science Education*, 31(10), 1379 - 1394.
- Kalender, I. (2009). Başarı ve yetenek kestiriminde yeni bir yaklaşım: bilgisayar ortamında bireyselleştirilmiş testler (computerized adaptive tests - CAT). *CITO Egitim Kuram ve Uygulama*, 5, 39-48.
- Berberoglu, G. & Kalender, I. (2007). Investigation of student achievement across years, school types and regions: the student selection examination and pisa analyses. *Educational Sciences and Practice,* 7(4).
- Kalender, I. (2006). *A structural equation modeling study: Comparison of factors affecting students' science achievement levels with respect to grade levels*. 7th National Science and Mathematics Education Congress, Turkey.
- Kalender, I. & Berberoglu, G. (2004). *A trend analysis of student selection examination*. 6th National Science and Mathematics Education Congress, Turkey.
- Kalender, I. (2004). *Use of computer adaptive tests in education*. 13th Educational Sciences Congress, Turkey.

Hobies

Probability in daily–life, comics, soundtrack collection.