COMPARABILITY OF SCORES FROM CAT AND PAPER AND PENCIL IMPLEMENTATIONS OF STUDENT SELECTION EXAMINATION TO HIGHER EDUCATION

A MASTER'S THESIS

BY

AYŞE SAYMAN AYHAN

THE PROGRAM OF CURRICULUM AND INSTRUCTION

İHSAN DOĞRAMACI BİLKENT UNIVERSITY

ANKARA

MAY 2015

Dedicated to my family with my love…

COMPARABILITY OF SCORES FROM CAT AND PAPER AND PENCIL IMPLEMENTATIONS OF STUDENT SELECTION EXAMINATION TO HIGHER EDUCATION

The Graduate School of Education

of

İhsan Doğramacı Bilkent University

by

Ayşe Sayman Ayhan

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

The Program of Curriculum and Instruction

İhsan Doğramacı Bilkent University

Ankara

May 2015

İHSAN DOĞRAMACI BİLKENT UNIVERSITY

GRADUATE SCHOOL OF EDUCATION

THESIS TITLE: COMPARABILITY OF SCORES FROM CAT AND PAPER AND PENCIL IMPLEMENTATIONS OF STUDENT SELECTION EXAMINATION TO HIGHER EDUCATION

AYŞE SAYMAN AYHAN

May 2015

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction.

…….

Asst. Prof. Dr. İlker Kalender

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction.

…….

Asst. Prof. Dr. Semirhan Gökçe

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Curriculum and Instruction.

……..

Prof. Dr. Alipaşa Ayas

Approval of the Graduate School of Education

……

Director Prof. Dr. M. K. Sands

ABSTRACT

COMPARABILITY OF SCORES FROM CAT AND PAPER AND PENCIL IMPLEMENTATIONS OF STUDENT SELECTION EXAMINATION TO HIGHER EDUCATION

Ayşe Sayman Ayhan

M.A., Program of Curriculum and Instruction

Supervisor: Asst. Prof. Dr. İlker Kalender

May 2015

The purpose of this study was to investigate the possibility of computerized adaptive testing (CAT) format as an alternative to the paper and pencil (P&P) test of the student selection examination (SSE) in Turkey. The scores obtained from both P&P format of the SSE and CAT through post-hoc simulations were compared using science subtest items. Different test termination rules (fixed length and fixed standard error) and ability estimation methods (EAP and MLE) were used to operate the CAT version of the SSE P&P test. 10, 15 and 25 items were used as fixed length test and standard errors of 0.30, 0.20 and 0.10 were used as fixed standard error thresholds in terms of test termination rules. Results indicated significant correlations between scores from SSE and CAT. The comparisons between results obtained from CAT and P&P tests also revealed that there exists similar ability distributions and

significant reduction in the number of items used through CAT. The findings from the research showed that CAT could calculate reliability using fewer items than P&P test. This study suggests that CAT can be an alternative to SSE with comparable scores to P&P format.

Key words: CAT, computerized adaptive testing, science achievement, student selection

ÖZET

YÜKSEK ÖĞRENİME GİRİŞ SINAVININ BİLGİSAYAR ORTAMINDA BİREYSELLEŞTİRİLMİŞ TEST VE KAĞIT KALEM TESTİ FORMATLARINDAN ELDE EDİLEN PUANLARININ KARŞILAŞTIRILMASI

Ayşe Sayman Ayhan

Yüksek Lisans, Eğitim Programları ve Öğretim

Tez Yöneticisi: YardımcıDoçent Doktor İlker Kalender

Mayıs 2015

Çalışmanın amacı yüksek öğrenime giriş sınavında bilgisayar ortamında bireyselleştirilmiş testin (CAT) öğrenci seçme sınavı (ÖSS) klasik kağıt ve kalem testlerine alternatif olabilirliğini araştırmaktır. Bu bağlamda öğrenci seçme sınavına ait fen alt testi kullanılarak hem kağıt ve kalem hem de CAT simülasyonlarından elde edilen puanlar kıyaslanmıştır. ÖSS sınavını CAT formatında yapılandırmak için sabit soru sayısı ve standart hata eşik değerleri ile farklı yetenek kestirim metotları (EAP ve MLE) gibi farklı test sonlandırma kuralları kullanılmıştır.Farklı yetenek kestirim metotları altında sabit soru sayısı değerleri 10, 15 ve 25; standart hata eşik değerleri 0.30, 0.20 ve 0.10 test sonlandırma kuralı olarak kullanılmıştır. Bu bağlamda ÖSS ve CAT simülasyon sonuçları arasında önemli bir korelasyon bulunmuştur. Ayrıca CAT ile soru sayısında önemli miktarda azalma ile benzer yetenek düzeyleri tespit edilmiştir. Bu çalışma sonucunda bireyselleştirilmiş testin daha az soruyla daha

güvenilir bir sınav sağladığı tespit edilmiştir. Bu sebepten çalışmaya konu olan araştırma bireyselleştirilmiş testi kıyaslanabilir skorlarla ÖSS kâğıt kalem testine alternatif olarak önermektedir.

Anahtar kelimeler: bilgisayarda bireyselleştirilmiş test, CAT, öğrenci seçme, fen başarısı

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Asst. Prof. Dr. İlker Kalender for his enthusiasm, encouragement and his resolute dedication to the strangeness of my knowledge on computerized adaptive tests. Also I am thankful to him for his criticism and detailed comments.

I am also indebted to my committee members, Prof. Dr. Alipaşa Ayas and Asst. Prof. Dr. Semirhan Gökçe for their feedback and criticism.

I take this opportunity to express my gratitude to the members of Graduate School of Education for their help and guidance.

Last but not least, I would like thank Annelise Hein and Andrew Bonar for proofreading my thesis and I am grateful to my husband who supported me throughout this venture.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAT	Computerized Adaptive Testing
CTT	Classical Test Theory
EAP	Expected A Posteriori
IRT	Item Response Theory
MLE	Maximum Likelihood Estimation
ÖSYM	Student Selection and Placement Center
P&P	Paper and pencil test
SE	Standard Error
SSE	Student Selection Examination

CHAPTER 1: INTRODUCTION

Introduction

The purpose of testing in education can be described as determining students' abilities/skills, providing feedback for instructors and students, school accountability and measuring pre-determined skills. Paper and pencil (P&P) testing format has been very common for decades in educational settings, but the wide use of computers may replace P&P format and provide significant improvement in testing, especially large scale testing which has been given in P&P format in Turkey. For example, Entrance Examination for Graduate Studies (ALES) and Student Selection Examination (SSE) are still done based on P&P format. Recently there have been some changes observed in testing format. The Foreign Language Examination for Civil Servants (YDS) test has been delivered via computer after 2014. The test only replaced the medium of delivery (from paper to computer), not the testing approach. Students still needed to proceed item by item as in a P&P test but it was the first time that computers were used for testing in Turkey in a large scale assessment.

Computer-based tests have been used for many years but most of them are linear tests. Since 1980s, a new testing format, computerized adaptive testing (CAT) has been proposed to change not only the testing medium, but also the testing philosophy. Instead of presenting items to all examinees in front of a computer, CAT dynamically selects and delivers items based on students' progress in the test. This study focused on the (i) applicability of CAT format to SSE, probably the most important large scale test, and (ii) comparability of scores obtained from conventional testing environment (P&P) and CAT formats.

Background

Student selection examination in Turkey for undergraduate programs

There are many examples for large scale testing implementations in Turkey. SSE is probably one of the most important ones since the scores from SSE are used for placement into higher education. SSE consisted of two phases in 2014. The first phase was Entrance Exam to Higher Education (YGS) which is used for placing students in some higher education programs and used for selecting students for the second phase. The first phase testing occasion includes 160 items. The subjects are Turkish, mathematics, social sciences (history, geography, philosophy and religion) and science (physics, chemistry, biology). Each subject group contains 40 items. The second phase is the Placement Exam to Higher Education (LYS) and used for placement (ÖSYM, 2014).

This exam unfortunately generates tremendous pressure on students; may even affect their psychology in negative ways (Yıldırım & Ergene, 2007). An exam administered once a year surely affects students' lives since their futures depend on this exam. If a student has health problems or cannot participate to exam, they have to wait until next year. There is an important detail here; one year may change everything in a human`s life.

In addition to the following problems, a psychometric problem should also be discussed in SSE. The results obtained from SSE across the years points to an important issue. For example, means of SSE 2005 show that the mean of science subtest is 3.9 out of 45. In the years 2008 and 2009, the mean of the 30-item science subtest was 3.9 and 4.0 for the students at Grade 12 (at the last year of the high school), respectively (ÖSYM, 2005; 2008; 2009). More recent means of the science

subtest from the SSE 2014 (4.5) and 2015 (4.6) indicated that means have been quite low (the means reported in this paragraph were calculated for all students who take the SSE in respective years, ÖSYM, 2014; 2015).

It is obvious that something is wrong with the balance between difficulties of items and the ability levels of examinees. The results for the last ten years show that these items did not match with the ability level of the students. The mismatch between difficulty level and ability level result in too much weight being given to items, rather than students' ability. In other words, a correct item in SSE may significantly change a student's rank. Another issue regarding the mismatch is that most of the items were not correctly answered by students. This situation may develop a claim that there may be no reason to ask many questions if they have a very low mean. Within these tests, some questions are difficult to solve for individuals having low ability and by the same way some are very easy for the students who are high achievers so it is meaningless to assess each individual with the same set of items.

In addition, regarding P&P format of SSE, there are other issues to consider such as security problems, transportation of booklets into exam centers, and organization of the exam for more than one million examinees for all over the country on a particular day. Moreover, the items are prepared for the P&P format which may be quite tiring or boring to read. Also they are not items designed by multimedia such as videos or animations which can be creative and may help to measure critical thinking skills of examinees (Çıkrıkçı-Demirtaşlı, 2003; Kalender, 2012).

Based on the problems across years, it can be said that a new test design is needed urgently which provides suitable items for examinee's ability level. If the test design maintains suitable items for each examinee, much more compact exam can be given

without delivering items too far above or below students' levels (Kalender, 2011). Also many issues regarding logistics such as time, security and transportation will be solved.

Computerized adaptive testing (CAT)

The idea of adaptive testing first appeared with the Binet& Simon intelligence test (Weiss, 1982). But "computerized" adaptive testing idea first appeared in the 1970s from the U.S. Department of Defense which highlighted the benefits of CAT (Wainer, 1993). Early studies were done by US Army, Navy and Air Force but they were not very successful. The Armed Services Vocational Aptitude Battery (CAT-ASVAB) was administrated in CAT format in the 1980s in the United States. Then the progress in computer technology led to the development in CAT applications. A notable example of CAT is the Graduate Record Examinations (GRE) which has been developed by Education Testing Service (ETS). The GRE was first done as CAT in 1993 (Schaeffer, Steffen & Durso, 1995) and the Test of English as a Foreign Language (TOEFL) was first given as CAT in 1998 in the United States. Today, TOEFL has been using internet-based non-adaptive test format (Kim & Huynh, 2007). Moreover, the Graduate Management Admission Test (GMAT) and the National Council Licensure Examination (NCLEX/CAT) developed CAT formats to get the license in nursing in the United States (Gokce, 2012). MATHCAT is another computer adaptive testing system developed by the National Institute for Educational Measurement (Cito) in the Netherlands (Verschoor & Straetmans, 2010).

CAT differs from P&P tests in many ways. First of all, it ensures that each student is given a test which includes items tailored for his or her ability level. Items are

dynamically selected from a large item pool based on the current ability levels of students, estimated immediately after each response.

In the background of CAT is Item Response Theory (IRT), a mathematical theory which relates ability level of students and item parameters. (Lee, Park & Kim, 2012). IRT based CATs produce comparable scores for test-takers who took the same test, giving them different weights. Another major advantage of CAT is the ability of estimating a reliability level of each student's score. In conventional testing, a common reliability is estimated for all examines.

Some advantages of CATs are given below (Tonidental, 2001; Zickar, 1998):

- Test security is improved since each examinee takes different items,
- Fast scoring can be provided just after the test,
- The possibility of cheating and copying is eliminated,
- Flexibility of administration is increased,
- Testing time is reduced,
- The number of items are reduced,
- Efficiency and measurement success are provided,
- Measurement precision becomes the same for each individual,
- Test is standardized,
- Items can be interactive such as animations.

On the other hand, there are some disadvantages of CAT, for example the cost of software programs, psychometric properties of the test, or requirements for a large item pool. In addition, CAT does not allow skipping the items. P&P tests allow examinees to go back and review the answers but CAT requires the answers to be given for an item in order to see the next item. Another disadvantage is that CAT and P&P tests differ in terms of scoring (Stocking, 1987). Conventional tests, based on classical test theory, calculate only correct responses while adaptive tests use IRT in that responses to prior items determine both the ability estimations of individuals and the selection of new coming items.

Problem

Statistics show that 1,282,512 Turkish middle school students took the Transition from Middle to Secondary Education (TEOG) exam in Grade 8, in 2015, and more than one million young people take the SSE each year (MoNE, 2015; ÖSYM, 2015).Especially SSE has significance since it is a high-stake test, the results of which are used for placement purposes in higher education institutions. The SSE has been given in P&P format for years.

The SSE consists of only multiple-choice items aiming to measure students' knowledge. Although the questions are developed based on the high school curriculum, they still do not match the ability of test takers (Berberoglu, 2012). Also, blind guessing is an important issue to consider about SSE. According toKalender (2012), "Giving a correct response to any item could change examinees' ordering significantly. When 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" (p.6).

Whether it is P&P or CAT, a test should be able to determine if the true score comes by blind-guessing or from the examinee's knowledge. According to Mead & Drasgow (1993), when the questions fit the ability of the examinee, there will not be blind-guessing problems. However, CAT may provide efficient solutions for all of those problems. First, CAT can provide items which fit well with the ability levels of examinees. Second, there will not be security or transportation of booklets thanks to the structure of CAT. Third, each examinee is given different items based on their ability levels which make cheating impossible. Fourth, items can be interactive and media presented which make the exam friendly. These conditions make the test more creative and lead to ask critical thinking questions or 3D items. Fifth, scoring will be easy right after the test. Also the most important thing is CAT could reduce the pressure on students if it is done more than once a year. To be able to solve all these problems, this study investigated if CAT can be an alternative to P&P based SSE.

Although the advantages of CAT format in large-scale testing programs are many, studies in Turkey are very limited (Demirtaslı-Cıkrıkcı, 2003; Gökce, 2012; Iseri, 2002; Kalender, 2011and Ozbası, 2014). These studies mostly focused on correlations between scores obtained from P&P and CAT formats. If testing medium may be given as an option to students, then students may select testing format. When this is the case, another problem arises. Scores from both formats should be comparable.

Despite the fact that CATs have become popular in other countries, it may not be easy to put SSE in CAT format immediately in Turkey. If two versions co-exist together, their scores should be comparable. There are four aspects of comparability which are reliability, test length, item exposure and content balancing (Wan, Ken,

Davis & McClarty, 2009). This study focused on comparability by using different text lengths and reliability.

Purpose

This study investigated if (i) CAT could be an alternative to SSE or not and (ii) how the scores obtained from CAT and P&P versions of SSE are comparable. Different test termination rules and ability estimation methods were used to operate CAT versions of the SSEP&P test. The results obtained from both P&P format of the SSE and CAT through simulations were compared. For this reason, SSE2005 science items were used to generate new results via post-hoc simulations. Different school types were also considered in this study. The results of general, Anatolian and private high schools were examined since these schools represent different ability levels (Kalender, 2011).

Research questions

The first research question is related to applicability of CAT. The sub-research questions were stated based on the studies in the literature on comparability issues between CAT and P&P tests (Wang & Kolen, 2001; Wang & Shin, 2010).

Could student's scores from P&P and CAT formats be compared?

1. Is there any reduction in the number of items required by CAT?
2. Is there a correlation between ability estimates obtained from CAT and P&P tests?
3. Is there any difference in difficulty between the CAT and P&P tests?
4. Are there any differences in terms of score distributions obtained from CAT and P&P tests?

5. Are there any differences in terms of the reliability of scores obtained from CAT and P&P tests?
6. For what percentage of test-takers is MLE not able to produce scores?

Significance

This study focuses on the viability of a reform to SSE with CAT as an alternative to the P&P of SSE, not only considering applicability of CAT but also comparability of CAT to P&P format. Transforming a P&P test into CAT format can easily be done. But if the comparability between different testing formats is not possible, then different formats cannot be delivered optionally. The results are expected to yield significant information especially for score comparability which may be used by Student Selection and Placement Center (ÖSYM), test developers, and measurement specialists.

Definition of terms

CAT (Computerized Adaptive Testing): A testing method based on IRT which tailors abilities of test takers according to their previous responses (Weiss, 1982).

CTT (Classical Test Theory): A measurement framework in which most of the P&P tests are grounded (Fan, 1998).

EAP (Expected A Posteriori): One of the ability estimation methods which belongs to Bayes' theorems (Boyd, 2003).

IRT (Item Response Theory): The theory in which CAT is grounded; it gives information about the ability of examinee on item level (van der Linden, 2000).

MLE (Maximum Likelihood Estimation): One of the ability estimation methods that does estimates based on the model parameters (Keller, 2000).

SSE (Student Selection Exam): The national exam for higher education in Turkey. In 2005, there were two phases of the university entrance exam. The first phase of the exam was called ÖSS (Student Selection Exam) and the second phase was called ÖYS (Student Placement Exam). Now, as of 2015, the first phase of the exam is called YGS (Entrance Exam to Higher Education) and the second phase is called LYS (Placement Exam to Higher Education). This study focused the first phase of the exam in 2005. The abbreviation of SSE is thus used to represent ÖSS.

SE (Standard Error): Standard error of ability estimation which is mainly used here as a test termination rule to assess reliability.

TEOG (Transition from Middle to Secondary Education): A test for Grade 8 students in Turkey. The results are used for placement in secondary education (MoNE, 2015).

*YDS (*Foreign Language Examination for Civil Servants*)*: The test to measure foreign language skills of candidates for placement in graduate programs in Turkey (ÖSYM, 2015).

CHAPTER 2: REVIEW OF RELATED LITERATURE

Introduction

Innovations in computers have significant effects on many areas of education, such as curriculum, measurement and evaluation. Previously, access to computers was limited to people from different levels of society because of high costs and capability. Since the early 1990s, it has been reported that this influx of technology has had an effect on student learning (Christensen, 2002). At the same time, CAT has been developed and implemented in large-scale testing programs such as licensure, certification, admissions, and psychological tests, especially in the United States, China and India (Kim & Huynh, 2007).

The purpose of this chapter is to provide context about large scale testing and problems. The definition of CAT, its advantages, the theory behind it, ability estimation methods and test termination rules are given to explain how CAT could be an alternative to SSE.

Large scale testing

Large scale testing in the world

Bennet (1998) states that large scale educational assessments have multiple purposes for a sizeable number of people such as placement, course credits, graduation or school accountability. GRE (Graduate Record Examination), TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System), SAT (Scholastic Assessment Test), IB-DP (International Baccalaureate Diploma Program), TIMSS (Trends in International Mathematics and Science

Study), and IAEP (International Assessment of Educational Progress) are some of the examples of these large scale testing from several countries (Eignor, 1993).

Large scale testing in Turkey

Turkey has a population of 76,667,864 people and 12,691,746 of them are between the ages of 15 and 24 (TUIK, 2013). Thus 16.6 % of the population consists of young people many of whom take at least one of the national exams, which are Public Personnel Selection Exam (KPSS), Student Selection Examination (SSE), Entrance Examination for Graduate Studies (ALES), Foreign Language Examination for Civil Servants (YDS) and Transition from Middle to Secondary Education (TEOG) Examination each year. For instance, 1,987,488young people took the SSE in 2015 and people of 1,783,313 took the KPSS in 2014at graduate level (ÖSYM, 2014; 2015).

Tindal & Haladyna (2002) stated that large scale assessments have many issues to consider and research which can be listed as follows: educational reforms, application of learning theories to standardized tests, validity, measurement and evaluation. Because of the fact that Turkey has many large scale tests, there are many issues to take into consideration.

Problems of large scale testing in Turkey

Recent results of SSE showed that the means of the science subtests were too low. Hence the mismatch between ability levels of students and item difficulty is probably the most important psychometric problem. If this is the case, a test may not be assessing what it is intended to measure. The security and transportation of questions are some of the problems about large scale exams. In 1999, one of the booklets of the exam was stolen and the exam was delayed. It cost the government a large sum of

money to repeat the test and prepare new booklets. While CAT may require some security issues, they are not more than a P&P test's.

The university entrance exam is very important in Turkey, as demonstrated by Berberoglu (2012). Millions of young people are trying to pass the exam and go on to higher education, but the reality is there are not enough places for each student. This causes a big competition among students, and also their families, who pay a huge amount of money for tutors or special courses. This puts pressure on students. According to Yildirim (2004), there is a significant correlation between depression, test anxiety and daily hassles of students. That means there is a need to reform of SSE.

Berberoglu (2012) states that a reform in SSE must be based on academic research and the research must examine two different issues: logistics and psychometrics. Administering the university entrance system only once in a year is harsh and hard for students. The exam needs to be rational and well-structured. The students' needs and aspirations, as well as their high school scores, must be taken into consideration. Also, multiple choice testing is a problem because the measurement is based on only test-taking ability. The multiple-choice-based education system of Turkey had low PISA scores in 2012 (OECD, 2012). It is clear that this needs to be changed and this change in education needs experts, reform must be based on scientific research, and the techniques must be suitable for measuring students' abilities.

CAT administrations

It was in early 1905s when Binet invented an adaptive test with the aim of asking questions to children to determine their intelligence level for their age groups (Weiss, 1982). The test was tailored because the difficulty of questions changed according to

previous answers that came from test takers. Some administrations of CAT are CAT-ASVAB (Armed Services Vocational Aptitude Battery) to measure academic and occupational success in military, CAT-GMAT for graduate business schools, Microsoft© Certified Professional Exams for certification in technology, American Institute of Certified Public Accountants Exam (AICPA) to become Certified Public Accountants, and NREMT (Nationally Certified Emergency Medical Technician) to become medical technician in United States(Fetzer, Dainis & Lamber, 2011).

CAT is a technologic assessment system, developed for computers. CAT involves change in both the administration mode and the test delivery algorithm, which turns CAT from linear to adaptive form. This algorithm allows the program to tailor each test (Wang, Kuo, Tsai & Liao, 2012). Rudner (2012) defines CAT as a test in which examinees are posed questions that will adjust in accordance to their responses to easy questions or difficult questions. To do this, CATs need a huge question databank.

Limitations of CAT

According to some comparability studies (Wang& Shin, 2010; Wang & Kolen, 2001), it has been found that test paradigm may be a factor for incomparability. Likewise, mode effect, which can affect examinees' performances, is one of the most important concerns about CAT. Ayberk (2012), Drasgow & Chuah (2006) and Schaeffer, Reese, Steffen, McKinley & Mills (1993) found that there is no significant difference between computer based tests and P&P test results in terms of gender and computer familiarity. According to Clark (1994), if the tests have the same content and cognitive activity, the results must be the same for both computer and P&P based tests. On the other hand, Clark (1994) stated that there is a significant difference

between student's computer-based test performances according to attitude towards computers. The cost of software programs, inability to review the test or skip items, and the need for many computers to run CAT are some of the disadvantages of CAT. In addition, item exposure, item selection and requirement for large item pools have significance to avoid repetition of item usage (Wainer, 1993).

This study focused on the comparability of the scores from P&P of SSE. However, there is very little research about CAT in Turkey and most of these studies are PhD dissertations. That is why this master thesis has significance for researchers who would like to conduct a further research about CAT in Turkey.

Item response theory

There are two widely accepted kinds of measurement frameworks, Classical Test Theory (CTT) and IRT. CTT does not have a complex theoretical model. As stated in Fan (1998); "CTT collectively considers a pool of examinees and empirically examines their success rate on an item" (p.358). On the other hand, IRT has a well defined theoretical model and it gives information about the ability of examinees on an item level. This property of IRT is used for adaptive testing. IRT is mentioned by Lord & Novick (1968) for the first time. As an example for linear test based on IRT, the Test of English as a Foreign Language (TOEFL) has been done all over the world for many years. It is a good example of computer-based tests. Many other standardized achievement and aptitude tests such as the Scholastic Aptitude Test, the California Achievement Tests, the Stanford Achievement Tests, and the Woodcock-Johnson Psycho-Educational Battery are developed using item response model principles and procedures (Hambleton, Zaar & Pieters, 1991).

IRT is the theory in which CAT is grounded. According to Hambleton, Zaar & Pieters (1991); "Classical test theory models and methods which have been in wide use for 60 years or more are being replaced by new test theories and methods, most notably item response theory" (p.341).IRT is a good theoretical basin for a test and provides useful information to test developers. In contrast to CTT, IRT gives information on item basis. It calculates the ability of the examinee on an item level by calculating the probability of the correct response for each item. It is not easy to assume examinees' response on an item basis but IRT makes it known. This is the main feature of IRT which means IRT calculates the probability of correct response for next item. For this reason, IRT can be called *probabilistic test theory*. In addition, this feature of IRT provides item characteristics independent from the group, ability estimations independent from the items and reliability estimations on individual levels. IRT derives item characteristics independent from the group, which means that item parameters would be same for individuals. This is called invariance of ability parameters. In addition, ability estimations are independent from the items which mean if two different sets of items are given to test takers; the same ability levels are estimated. This is called invariance of item parameters. To be able to generate item parameters, items are calibrated based on their difficulty levels: the items are given to test takers to generate item parameters based on their responses. When the model and item parameters fit, the test can be given to different examinees with different items to receive comparable estimates of ability. Therefore, the items in the test match with the ability levels of individuals (Embretson & Reise, 2000; Hambleton & Jones, 1993; Kingbury & Zara, 1989; Mead & Meade, 2010; Reise & Waller, 2003; Stocking, Smith & Swanson, 2000; Yen, 1981).

IRT has several models that aim to calculate the probability of certain responses for a certain item. The models have functions which are normal ogive or logistic functions. The models in logistic functions are explained below (Hambleton & Swaminathan, 1984).

One parameter logistic model (1PL) is interested in item difficulty, *b*. This is the simplest dichotomous IRT model since there is only one item parameter. This item parameter, *b*, provides information about the person's ability level since there is an interaction between the difficulty of item and ability. Therefore, the probability of a correct response can be predicted.

One Parameter Logistic Model:

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \qquad i = 1, 2, \ldots n \qquad (D = 1.7)$$

Two Parameters Logistic Model:

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \qquad i = 1, 2, \ldots n \qquad (D = 1.7)$$

Three Parameters Logistic Model:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \qquad i = 1, 2, \ldots n \qquad (D = 1.7)$$

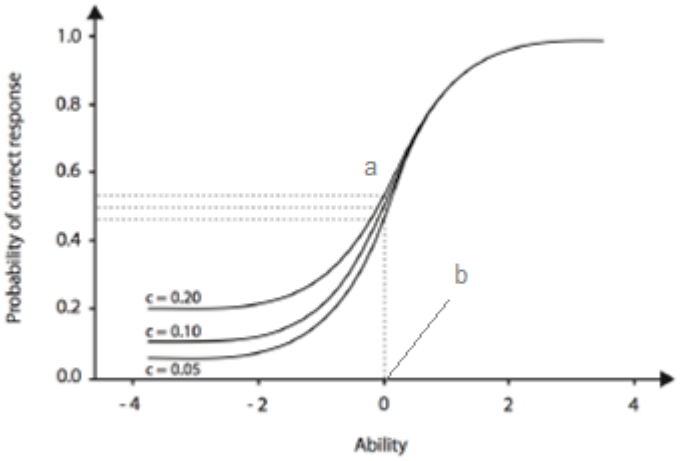


Figure 1.Three parameter logistic model item characteristic curve

This item characteristic curve shows the relationship between ability and probability of a correct response (P_θ) for an item. *c* represents the pseudo-guessing factor, *b* represents item difficulty and *a* represents discrimination. As *b* goes right, items become harder; and as *b* goes left, items become easier. As can be seen from the Figure 1, 0.5 is the changing point for the parameter that makes the graph flatter or steeper. The item characteristic curve gives information on item level.

Much of what IRT provides comes from *item information function*. The idea is to match the item difficulty with the ability level of examinee. Then the information is received with regards to how close the difficulty of the item is to the ability of examinee. The amount of information depends on this criterion. When the ability is known for each item, then the difficulty can be calculated. To be able to have a good measurement, it is important to have variety of items, which have different difficulty levels (Bock, 1972; Pashley, 1991; Rizopoulos, 2006).

The two parameters logistic model (2PL) is labeled as Birnbaum model whereas the one parameter logistic model (1PL) is called as Rasch model in literature. The 2 PL model uses two parameters which are item difficulty, *b*, and item discrimination, *a*.

These two parameters derive information when the difficulty of an item is too high for an examinee, who has low ability. Higher ability level is achieved by higher difficulty of items. Likewise, item discrimination distinguishes between the items with different ability levels (Kalender, 2009). The three parameter logistic model (3PL) has the additional influence of guessing parameter, *c*, on item difficulty, *b*, and item discrimination, *a* (De Ayala, 2009; Pashley,1991).

Some IRT studies from Turkey are given below:

Baykul (1979) investigated the results obtained from 3PL model and CTT in terms of mathematical test ability. In the first place the results stated that 3PL generated more reliable scores. Berberoglu (1988) studied the contributions of Rasch Model which is a special case of IRT; to operate measurement and to increase objectivity of test items of SSE. The results were compared to the results from CTT. The findings showed that the Rasch Model operated higher scores in terms of both reliability and validity.

The study which was done by Demirtaslı-Cıkrıkcı (2003) examined the comparison of Raven Standard Progressive Matrices (RSPM) Test under CTT with IRT models. The results revealed that item difficulty indices for both CTT and IRT were highly invariant, and IRT based CAT were suggested as a test application. In addition, Kalender (2011) investigated the effects of different CAT strategies on the recovery of ability. The results obtained from IRT and CTT were compared. SSE 2005, 2006 and 2007 science sub-test items were used to generate scores by CAT. The findings underlined high correlations between the scores from conventional P&P test and CAT. Lastly Iseri (2002), Kaptan (1993), and Ozbaşı (2014) did research about IRT in their dissertations and explained many advantages of IRT over CTT.

Fan (1998), Samejima (1969), Van der Linden & Glass (2000) and Zickar (1998) did an empirical study to identify differences between IRT and CTT since they are very different measurement frameworks. The findings supported the idea that IRT has many advantages over CTT.

Ability estimation methods

There are four ability estimation methods in literature (Beguin & Glas, 2001; Keller, 2000). As it is stated in van der Linden (2010), the most popular one was maximum likelihood estimator, MLE. The others are weighted likelihood estimation (WLE), expected a posteriori estimation (EAP) and maximum a posteriori estimation (MAP). The last two are called Bayesian ability estimation methods. According to Beguin & Glas (2001) "Bayesian approach gives the possibility to rigorously model all dependencies among variables and sources of uncertainty (p.541). Moreover, Bayesian techniques provide flexibility and calculate ability for complex data (Fox, 2010). WLE can be an alternative to MLE since it uses likelihood estimations while the other two use posterior based estimations.

MLE focuses on maximizing likelihood to estimate ability. Lord (1986) discussed the advantages and disadvantages of MLE in IRT. Lord stated that MLE produces non-convergence results, which can be a disadvantage, whereas it has consistency and efficiency, which can be stated as advantages. Wang (1997) examined MLE and EAP ability estimates in CAT. It is concluded that MLE produces unbiased ability estimates and low standard error values whereas EAP produces biased results but high standard error values. The most important characteristics of EAP can be counted as there are not any non-estimated scores whereas non-convergence is a big issue for MLE. MLE requires at least one correct and one incorrect answer for scoring. If all

the answers are correct or incorrect for the set of items, MLE does not work, according to Hambleton & Swaminathan (1984).

The current literature on ability estimates abounds with examples of ability estimation methods. Yi, Wang & Ban (2000) did comparison of four ability estimation methods and WLE was found to be the best ability estimation method for ACT Mathematical Test. Likewise, Riley and Carle (2012) and Veldkamp & Matteucci (2013) examined the advantages and limitations of Bayesian CAT. McBride (1977) discussed some properties of Bayesian methods in adaptive testing. The results stated that Bayesian techniques highly correlated with ability levels. Chen & Choi (2009) discussed the difference between MLE and EAP. The findings reported that MLE produced non-convergence scores whereas EAP did not. In addition, MLE did not work for small sample sizes whereas EAP worked well. It is stated that "EAP estimators with more informative prior distribution could result in stronger bias towards the mean of the prior distribution, and provide less variation of estimates in terms of standard deviation" (p. 352). Lastly, Wang et al. (2012) investigated adaptive systems for Chinese proficiency. Results supported the key argument that EAP had many advantages over MLE and MAP.

Strategies for test termination

To end a CAT there are several methods. Fixed length CAT, variable-length CAT, passing scores, cut points, and standard error thresholds are the most common test termination rules in the literature (Wan et al., 2009; Wang & Kolen, 2001). A fixed set of items are given to test takers in fixed length CAT whereas different sets of items are used for variable-length CAT. Minimum standard error or a time limit is determined to stop variable-length CAT which fits with the minimum standard of

reliability. Passing scores or cut points can also be established as a test termination strategy for CAT (Wall &Waltz, 2004).

According to Weiss & Kingsbury (1984), there are important components to consider for CAT administrations. These are response mechanism, item pools, starting rule, item selection procedure, scoring models and test termination rules. The study that belongs to Weiss & Kingsbury (1984) has significance since different termination rules were applied with many item banks. Termination rule is important to catch efficiency in measurement. Otherwise, CAT cannot fit into good measurement tools. Babcock & Weiss (2012) used MLE as ability estimate method and reliability of 0.85, 0.90 and 0.95 for standard error thresholds together with other stopping conditions. Babcock & Weiss ‘s results showed that fixed length and variable-length CATs performed similarly whereas Boyd’s (2003) investigation showed that fixed length tests were more useful to calculate item exposure rate.

Summary

In this chapter, the discussion pointed out many articles and research papers to answer research questions of this study properly. In the first place, the concept of large scale testing was identified. The situation of large scale testing all over the world and in Turkey was detailed with many examples. Nearly all of the large scale testing was P&P based whereas there were only several tests that were computer based. Then, SSE of Turkey was highlighted since it was one of the most important exams in Turkey. The problems with large scale testing were examined. Transportation, security issues, and anxiety of once a year exams were the main topics to discuss regarding the problems of large scale testing in Turkey. After all, the idea of adaptive test was explained.

Next, the definition of CAT was given to express information about it. Advantages of CAT were explained to underline the comparability of CAT with P&P of SSE. Moreover, the theory behind it was detailed. The comparisons between IRT and CTT were given and on the basis of currently available evidences from the literature, IRT were highlighted with many advantages. In this study, EAP and MLE were used to examine different results and to find the optimum CAT strategy. For this reason, different ability estimation methods were detailed under the light of literature. Different test termination rules were given from the literature and the main theoretical premises behind test termination rules were detailed.

To conclude, in this chapter, the situation of P&P based large scale exams and CAT as an alternative to P&P tests was emphasized. CAT was presented as a good alternative to P&P tests. Since there is little research in Turkey about CAT, this study fills an important gap for future test developments.

CHAPTER 3: METHOD

Introduction

This chapter explains the methodology of the thesis. This study aims to investigate the applicability of CAT as an alternative to P&P testing with regard to the SSE in Turkey. In this section, there are six main parts. In the first section, research design was described. Then in the second section information context and in the third section sampling of the study were given. After that in the fourth section instrumentation of the study was provided. In the fifth section method of data collection and lastly in the sixth section the method of data analysis were detailed.

Research design

This study utilized quantitative methodology based on simulations. Descriptive research design was used to investigate results obtained from real data and simulations. Results were presented in a descriptive manner rather than conducting statistical comparison analyses since this is one of the earliest attempts regarding CAT implementations.

Context

In this study, students' responses to the science subtest of SSE were used.SSE2005 science items were obtained via official permission of the ÖSYM in 2005 without any ID or information that can be used to identify students. The present study includes different methods of comparability for SSE between CAT and the P&P versions. According to the literature, there are many methods to analyze comparability of CAT and P&P tests (Wan et al., 2009; Wang & Kolen, 2001). In this study, fixed length and fixed standard error methods were used together with

different ability estimation methods. Post-hoc simulation was used for applying these methods and new data was generated based on 2244randomly chosen real examinee responses. Responses belong to the students who took the SSE 2005 at Grade 12 (the last year of high school). The SSE 2005 had two phases but this study focuses on only the first phase of the national exam. Science items were used for generating post-hoc simulation results. Based on the results from the simulation, the comparability of CAT and SSE P&P version would be investigated.

Sampling

The data used in this research is obtained from ÖSYM. The data set belongs to year 2005 and it contains student responses to 45 science items. The sample size was 2244 students. The sample included only students who took the SSE2005 at Grade 12 (the last year of the high school) and those who gave at least one response to the science items. Beyond all, three school types are included: general, Anatolian and private high schools. These schools were included to represent different ability levels based on the mean scores obtained from SSE 2005 science sub-test (Table 2). General high schools had the highest percentage in terms of number of students for the sample of the study. These three school types followed the same national curriculum. Vocational and technical high schools were not considered for this study because these schools have extremely low means. But the school types in this study have relatively higher means.

As can be seen from Table 1, the mean of the total scores out of 45 science items is 18.31 for the sample of this study. In chapter 1, the mean scores were given including all students who took SSE 2005.In this study, only general, Anatolian and private schools were included.

Table 1
Descriptive statistics of total science scores of SSE 2005 (N=2244)

	Mean	Standard Deviation	Skewness	Kurtosis
Total_scores	18.31	14.41	.05	-1.41

Figure 2 shows the distribution of total scores of SSE 2005. As can be seen from the figure, there are many students who had total scores that equal to 0. Table 2 shows that the median values of total scores for the 2005 science subtest are 7, 34.5 and 26 for general, Anatolian and private high schools, respectively. Anatolian schools have the highest total scores whereas general high schools have the lowest total scores.

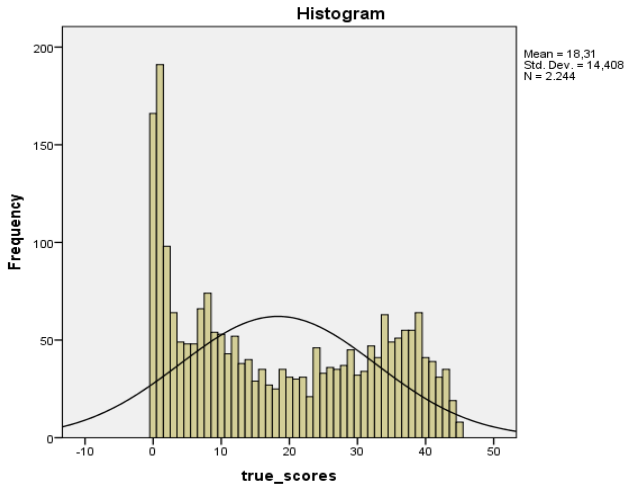


Figure 2. Distribution of total scoresof SSE 2005 (N=2244)

Table 2
Statistics of total scores of general, Anatolian and private high schools of SSE 2005 (N=2244)

	General	Anatolian	Private
Mean	9.95	31.41	23.96
Median	7.00	34.50	26.00
Mode	1.00	39.00	37.00
Std. Deviation	9.83	10.74	13.45
Variance	96.70	115.50	180.94
Skewness	1.10	-1.44	-0.34
Kurtosis	0.50	1.47	-1.12
Range	44.00	45.00	45.00

In Figure 3, it can be seen that general high schools have many students who have 0 out of 45 items. On the other hand, Anatolian school students have total scores around 30to 45. Evidently they represent the highest achievers.

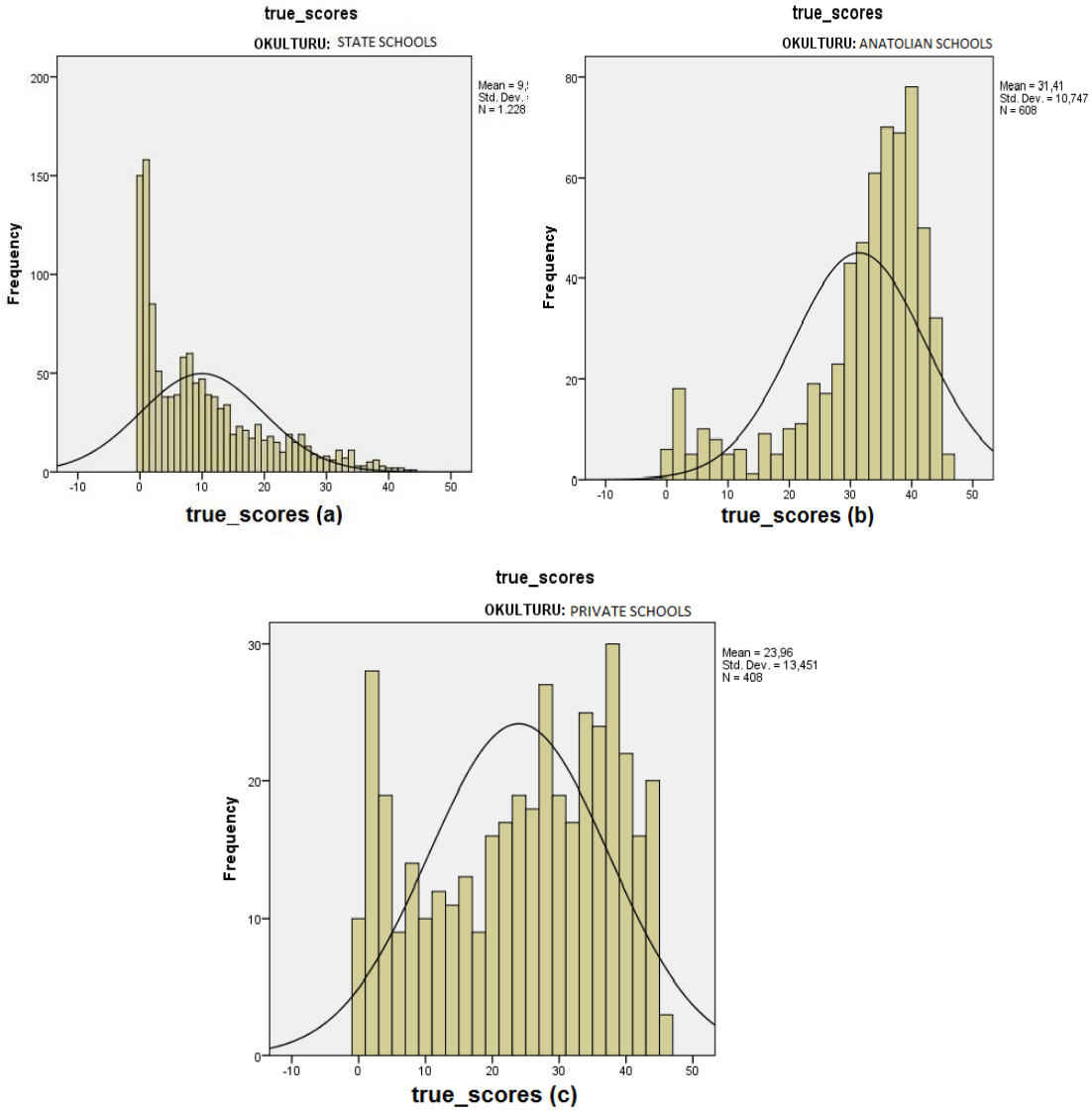


Figure 3. Distribution of total scores of general (a), Anatolian (b) and private (c) schools(N=2244)

In Table 3, it can be seen that the percentages of true responses for each item level are 63.4 maximum and 15.7 minimum. Missing items are the ones that were left blank by the examinees. The maximum percentage of missing items equals 52.5, which is very close to maximum value for the percentages of true responses on item

level. On these grounds, it can be argued that science items are very difficult to answer. This leads to an important discussion regarding the mismatch of the ability levels of students and difficulty of items. Most of the students do not prefer blind-guessing since four wrong answers erase one of the correct responses. Students prefer not to answer and leave the items blank. This situation creates a high percentage of missing values and low percentage of total scores. To summarize there are notable differences between the difficulty of the science items and the ability of test takers.

Table 3
Percentages of true, false and missing of 45 science items

Question numbers	True	False	Missing	Question numbers	True	False	Missing
1	57.0	16.2	26.8	23	30.7	16.8	52.5
2	15.7	51.4	32.9	24	44.5	11.9	43.6
3	54.7	17.1	28.2	25	29.2	28.9	41.9
4	62.1	14.3	23.6	26	26.0	41.7	32.3
5	39.0	35.3	25.7	27	39.3	16.0	44.7
6	27.1	30.8	42.1	28	22.0	36.5	41.5
7	47.5	23.2	29.3	29	30.5	23.8	45.7
8	44.6	20.2	35.2	30	45.5	16.4	38.1
9	48.1	25.3	26.6	31	48.6	18.4	33.0
10	22.3	46.6	31.1	32	46.3	20.6	33.1
11	40.9	37.7	21.4	33	42.8	19.6	37.6
12	54.7	18.8	26.5	34	49.4	17.3	33.3
13	46.6	24.1	29.3	35	47.8	20.6	31.6
14	56.3	15.2	28.5	36	40.5	15.2	44.3
15	43.2	25.9	30.7	37	22.3	35.4	42.3
16	31.1	24.8	44.1	38	36.8	30.5	32.7
17	38.2	30.4	31.4	39	41.5	13.8	44.7
18	63.4	15.0	21.6	40	25.9	29.8	44.3
19	36.1	26.3	37.6	41	24.0	36.8	39.2
20	52.4	21.7	25.9	42	44.8	18.2	37.0
21	56.5	16.0	27.5	43	24.5	32.5	43.0
22	44.3	28.9	26.8	44	40.9	19.5	39.6
				45	45.6	16.4	38.0

Instrumentation

The instrument of this study is the science sub-test items, which belongs to the first phase of SSE 2005. There were 45 science items in the SSE 2005. It consisted of 19 Physics, 14 Chemistry and 12 Biology items. This study focused on only the first phase of the SSE 2005 which is used for selection. It is stated in the SSE booklets that the aim of the test is to measure basic comprehension and thinking skills of students in science (ÖSYM, 2005).

Method of data collection

No data were collected for this study. Data sets including students' responses to science subtest were provided by the ÖSYM.

Method of data analysis

Since the CAT format requires item parameters defined in IRT, first, data sets were calibrated to obtain parameters for each item with respect to the three-parameter model (3PL): pseudo guessing, item difficulty and item discrimination. BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) program was used for calibration of the items. But before that data were converted to dichotomous format: correct scores were coded as 1, while the wrong ones 0. After defining item parameters, a series of post-hoc simulations were conducted. In these simulations, a testing environment was simulated as if students were given a CAT test, using their responses they gave earlier for the P&P test.

Post-hoc or real data simulations consider real examinees responses that have been administered conventionally. The aim is to reduce the number of items given by SSE 2005 P&P exam. In this case the item pool was generated by the same items with

SSE 2005 science sub-test and the data which was the responses of real examinees were analyzed. The software for post-hoc simulation was developed by Kalender (2011).

The working principle of post-hoc simulation is as follows:

- When the simulation starts, the computer picks the item for the examinee and then checks the response of the item for the same examinee from the P&P data since the items were used before in a conventional P&P test.
- Then the computer picks another item based on pre-determined item selection rules and checks the response of the examinee for that item.
- Items are chosen according to Maximum Information which means selected items have to gather the highest information.
- The computer does the same thing until it obtains pre-determined test termination rules. In this case it is fix length and fixed standard errors.
- After the simulation phase, several analyses were conducted to investigate the results.

First, numbers of items given to examinees under different post-hoc simulation were presented then the findings were interpreted. Next, correlations were calculated between the results obtained from CAT and P&P test based on different ability estimation methods and test stopping conditions. Ability levels for general, Anatolian and private schools were compared under fix length and fixed standard error test termination conditions. In addition, distributional features of all school types' ability estimates from CAT simulations and P&P test by MLE and EAP were presented. Then, distribution of standard error values for fixed items under the two ability estimation methods were given to compare results based on the school types. Last, the number of examinees whose scores were not calculated by MLE provided.

In this study, different school types (general, Anatolian and private) were investigated since they represent different ability groups. Two ability estimation methods were used for this study. They are MLE and EAP.

Different standard errors and different text lengths were used as test termination rules. By using scores from different administrations of CAT simulations, the comparability of SSE with CAT was identified. Recent studies showed that comparability of scores from different testing formats should be considered and checked by using appropriate methods (Vispoel, Rocklin, & Wang, 1994; Wang, 1997; Wang et al., 2007, 2008).

Test termination rules were used to end the test when enough information is gathered to estimate ability. Two different test termination rules were used to conduct this study. Fix length and fixed standard error (SE) were used as test termination rules; MLE and EAP were used as ability estimation methods. In fixed length tests, a fixed number of items is given to examinees. In this case, 10, 15 and 25 items were used to create the CAT format of SSE. In this way, different SEs were obtained. SE measures accuracy in a test. SE is the mean of the standard deviation of the sampling distribution. When SE increases, reliability decreases. For this purpose, SE 0.30, 0.20 and 0.10were used as fixed SE test termination rule. In CTT, these values correspond to 0.91, 0.96 and 0.99 reliability values. As a result, simulations were conducted by changing ability estimation methods (MLE and EAP) and test termination rules (SE=0.30, 0.20, 0.10 and fixed number of items: 23%, 33% and 55% of the P&P test) for each of the three school types.

Summary

This chapter consisted of six main parts, namely research design, context, sample, instrumentation, data collection and data analysis procedures. The first part provided

information about the type of research design used in the study and to find the possible answers to the research questions. The second part provided information about the context information of the study. The third part focused on the sampling strategy of this study; also gave detailed information about the school types, and students' scores. The fourth part, instrumentation, explained the tool used for the study. The fifth part focused on data collection methods. The sixth part examined how data were analyzed and reported for each research question.

CHAPTER 4: RESULTS

Introduction

This chapter gives information about the results of the current study. The results from different post-hoc CAT simulations and their comparisons to P&P tests are detailed here. Thus, research questions were analyzed sequentially. This chapter consists of six main sections. In the first section, important findings about reduction in the number of items by CAT were detailed compared in order to find the optimum strategy. The results under different test termination rules were given in order to find the optimum strategy (see Table 4). In the second section, ability estimations obtained from two different methods, EAP and MLE, by CAT were compared with the ability estimations obtained from SSE P&P2005 and investigated to see if there was a correlation between CAT and P&P of SSE in terms of ability estimations. In the third section, ability estimations obtained from CAT and SSE P&P 2005 test were compared to see the difficulty levels of the tests. In the fourth section, distribution of scores was presented according to the school types. In the fifth section, the fixed length test termination rule was applied to examine which one produced less SEs. Finally, the issue of non-convergence was analyzed by MLE and the findings were presented.

In this chapter, the results were given in the order of the research questions stated in Chapter 1.

Is there any reduction in the number of items required by CAT?

For standard error threshold-based post-hoc simulations, number of items was investigated to assess if there was any reduction, which was provided in test length

by CAT. The numbers of items given to examinees in post-hoc simulations under different CAT strategies can be seen in Table 4.

Table 4
The numbers of items given to examinees under different CAT strategies

	Ability Estimation Method					
	MLE			EAP		
	SE Threshold			SE Threshold		
School Type	0.30	0.20	0.10	0.30	0.20	0.10
General	5	7	45	13	26	45
Anatolian	8	23	45	15	29	45
Private	8	22	45	25	45	45

SE is important since it gives information about the reliability of a test. If SE decreases, the reliability of the test increases. As expected, number of items required to finish CAT increased with required level of SE decreased. For instance, SE 0.10 is equal to 0.99 in CTT. 0.99 represents the maximum reliability but it required 45 items in simulations. Evidently SE 0.10 is not working for tailored test simulations since it required the full length in P&P test, 45 items. For this reason, SE 0.10 was not used for further analysis.

Table 4 presented that SE threshold with 0.30 and 0.20 required 5 and 7 items administrated for general high schools. This means that general high schools by MLE estimates required fewer items than the others, but the number of items fewer than 10 may not be enough for a test due to the validity issues. On the other hand, EAP required more items than MLE for all SE thresholds hence SE with 0.30 and 0.20 results by EAP required 13 and 26 items for general high schools, respectively. SE 0.30 by MLE used 8 items whereas SE 0.20 by MLE used 23 items for Anatolian schools. Anatolian schools required more items in contrast to general high schools in all tested conditions. In fact, Anatolian schools contain successful students so those

schools were expected to have more reliable results than general or private high schools.

As seen in Table 4, SE 0.30 by EAP used 15 items for private schools whereas SE 0.20 administrated 29 items. SE 0.20 by MLE used 22 items whereas SE 0.30 required 8 items. Unfortunately SE 0.20 by EAP did not work for private schools since it required 45 items. SE 0.30 by EAP used 24 items which can be acceptable for CAT applications. At last, SE thresholds showed that for EAP and MLE ability estimations of SE 0.30 required fewer items than SE 0.10. In general Table 4 presented that EAP required more items than MLE and general high schools had the lowest mean across all tested conditions.

Is there a correlation between ability estimates obtained from CAT and P&P tests?

Table 5 shows the correlations between ability estimates obtained from CAT simulations and P&P test for all the different conditions based on different ability estimation methods and test termination rules. Simulation results of different conditions are MLE/EAP and SE thresholds /fixed item are given in Table 5.

A closer look at the data indicated that general high schools demonstrated lower correlation whereas private schools had the highest correlation for both ability estimation methods and test termination rules. The table yielded by this study provided convincing evidence that correlation between ability estimates by EAP method are much larger than MLE. In fact, MLE had the lowest correlation for general high schools. On the other hand, EAP had more stable results and showed higher correlation for both general high schools and other school types. According to the results given in Table 5, EAP, regardless of test termination rules, seemed to

work better in ability estimation. In addition, correlations between estimates for EAP were invariant for all conditions such as different school types and test terminations

Table 5
Correlations of ability estimates between CAT and P&P

	MLE					EAP				
	SE Threshold		Fixed Item			SE Threshold		Fixed Item		
	0.30	0.20	10	15	25	0.30	0.20	10	15	25
General	.71	.74	.75	.73	.80	.93	.95	.91	.94	.97
Anatolian	.88	.97	.91	.96	.98	.97	.98	.95	.96	.98
Private	.96	.99	.93	.96	.99	.98	.98	.96	.97	.98

The data generated by MLE was also reported in the Table 5. This table showed that MLE produced a variety of correlations for all tested conditions. Differences were visible especially for Anatolian schools under different SE thresholds. As seen in Table 5, when the number of items increased, correlations between estimates for both MLE and EAP also increased. These results developed the claim that EAP estimation method showed higher correlations which were more than 0.90 in all tested conditions. Therefore, EAP can be a better choice for ability estimation method in using the adaptive version of SSE.

Is there any difference in difficulty between the CAT and P&P tests?

Table 6 presents medians of ability distributions obtained under different post-hoc simulations (since the ability distributions were skewed, median was preferred). The data from post-hoc simulations showed that ability estimations for pre-determined SEs were invariant for both ability estimation methods. Fixed length test results presented that the EAP ability estimations may slightly differ from each other. Also MLE estimates were lower as compared to EAP estimates. It seems that Anatolian

and general high schools were provided slightly harder test by MLE than it was by the conventional P&P test.

Table 6
Median of ability estimates based on different post-hoc simulations

School Type	Ability Estimation Method	P&P	SE Threshold		Fixed Number of Items		
			0.30	0.20	10	15	25
General	MLE	.63	.65	.57	.57	.56	.61
Anatolian		.08	-.18	-.12	-.19	-.14	-.11
Private		.13	.04	.08	.10	.11	.10
General	EAP	.69	.38	.42	.28	.34	.42
Anatolian		.47	.35	.40	.22	.25	.34
Private		.38	.36	.41	.26	.24	.35

Table 6 showed that there was a difference between CAT and P&P ability estimations. Results indicated that CAT-based ability estimates were lower than those from P&P, indicating CAT delivered more difficult test to examinees. However, it should be noted that ÖSYM uses a different calculation method for the P&P format of SSE. Ability estimates are calculated based on correct and incorrect responses. The ÖSYM erases one true answer for four false answers in order to obtain more reliable scores but CAT uses different methods to calculate ability estimations.

Are there any differences in terms of score distributions obtained from CAT and P&P test?

Distribution of both CAT and P&P ability estimations by MLE obtained from general high schools

Table 7shows the distribution of both CAT and P&P ability estimations by MLE obtained from general high schools. The findings of the CAT simulations were relatively different from P&P ability estimations. On the other hand, both SE

thresholds and fixed number of items had stable results in CAT, which meant the findings obtained from CAT were similar internally. SE 0.20, fixed numbers of 15, and fixed numbers of 25 items had the same value for the mode which was, -3.00. SE with 0.30 by MLE seems to have produced the closest result to the P&P test (0.65, 0.63).

Table 7
Distribution of both CAT and P&P ability estimations by MLE obtained from general high schools

		SE thresholds		Fixed Number of Items		
	P&P	0.30	0.20	10	15	25
Mean	.02	.73	.55	.64	.52	.57
Median	.63	.65	.57	.57	.56	.61
Mode	-1.08	.40	-3.00	.37	-3.00	-3.00
Std. Deviation	1.04	.45	.81	.53	.86	.78
Skewness	.30	.11	-2.90	-1.26	-2.69	-2.91
Kurtosis	-.09	6.13	11.20	12.61	9.52	11.92

In Figure 4, the distribution of MLE P&P ability estimates for general high schools was given. The normal curve of the histogram was in leptokurtic form. It seemed from the Figure 4 that there were two mode values for general high schools by MLE P&P ability estimates. Approximately half of the ability estimates were above the mean, 0.02, and half of them were below 0.02, close to -3.00. This meant that general high schools had a distribution which was not normal in terms of ability estimates by MLE P&P. It could be read from the histogram that the mean value was 0.02 for MLE P&P ability estimates. Since the median was equal to 0.63, it was higher than the mean value for general high schools by MLE.

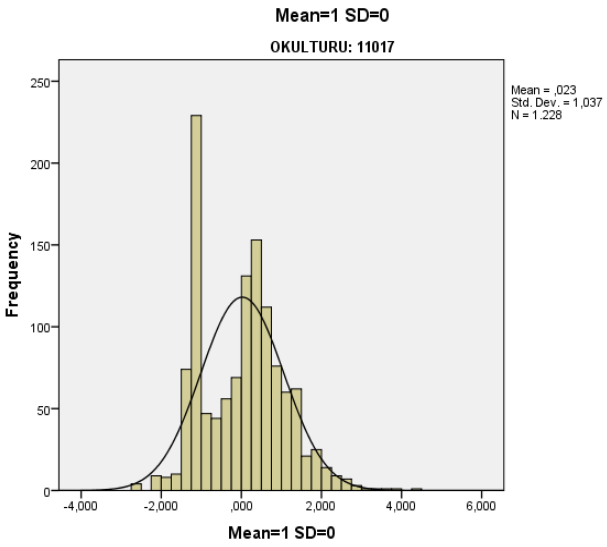


Figure 4. Distribution of MLE P&P ability estimates for general high schools

Figure 5 shows that the mean of SEs were equal to 0.41. SE scores of the simulees differed between 0.20 and 1.00 and there were only a few simulees who had SE scores more than 1.00 up to 2.50. Most of the simulees had SE value below 0.40 right-skewing the SE values. While SE 0.40 indicates a high reliability, it was expected that SE scores would be under 0.30 in this study. As seen from the Figure 5, MLE produced a higher range in terms of SE values, which at some parts decreased reliability.

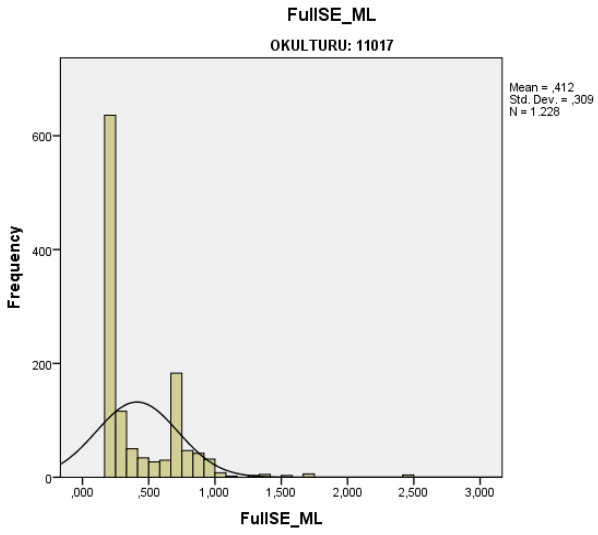


Figure 5. Distribution of standard errors by MLE P&P test for general high schools

Distribution of both CAT and P&P test ability estimations by EAP obtained from general high schools

A closer look at the data shown in Table 8 indicated that CAT produced lower ability estimates for general high schools. SE thresholds 0.30 and 0.20 had the same mode value, which was -1.30. The median of the ability estimations was higher than the mean value for P&P test. The mean of ability estimations was 0.02 for EAP P&P test whereas it was 0.23 by MLE P&P test. The median of ability estimations was 0.69 by EAP P&P test, whereas it was 0.63 by MLE P&P test for general high schools. Evidently, MLE produced lower ability estimations than EAP for general high schools.

Table 8
Distribution of both CAT and P&P test ability estimations by EAP obtained from general high schools

	P&P	SE Thresholds		Fixed number of items		
		0.30	0.20	10	15	25
Mean	.02	.13	.12	.09	.08	.09
Median	.69	.38	.42	.28	.34	.42
Mode	-1.39	-1.30	-1.30	-.98	-1.08	.44
Std. Deviation	1.01	.87	.87	.86	.88	.91
Skewness	-.03	.08	.10	.30	.26	.17
Kurtosis	-1.08	-.46	-.45	-.79	-.69	-.71

The histogram in Figure 6 shows a different ability distribution with two mode values. There were a high number of simulees who had ability estimations by EAP P&P between -1.10 to -1.50 and there were a high number of simulees who had ability estimations by EAP P&P between 0.10 and 1.40. As shown in the Figure 6, most of the simulees had ability estimates between 0.00 and 1.50. Therefore, it can be said that the results of EAP P&P ability estimations have a larger range than MLE P&P test.

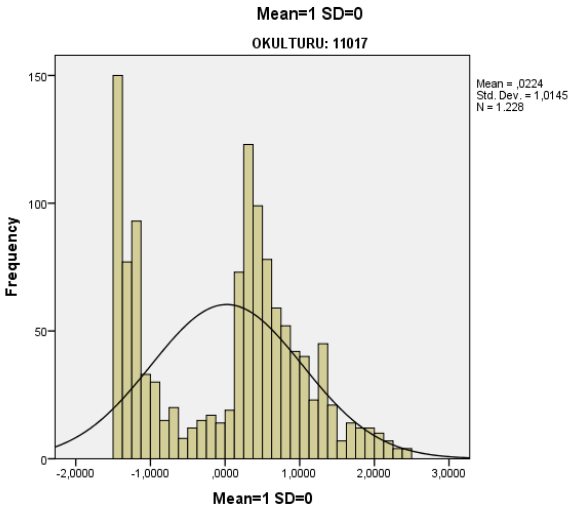


Figure 6.Distribution of EAP P&P ability estimates for general high schools

Figure 7 shows the distribution of SEs by EAP P&P test for general high schools. The mean of SE was 0.28. Most of the SE values were under 0.40, which makes the results reliable, and there were some simulees who had SE scores up to 0.80 for whom ability estimates may not be reliable. However, the histogram was right-skewed, just like MLE P&P test results, and the range of SE values was higher than MLE P&P test. There were a high number of simulees who had SE 0.10; a significant number of the simulees had SE above 0.40.

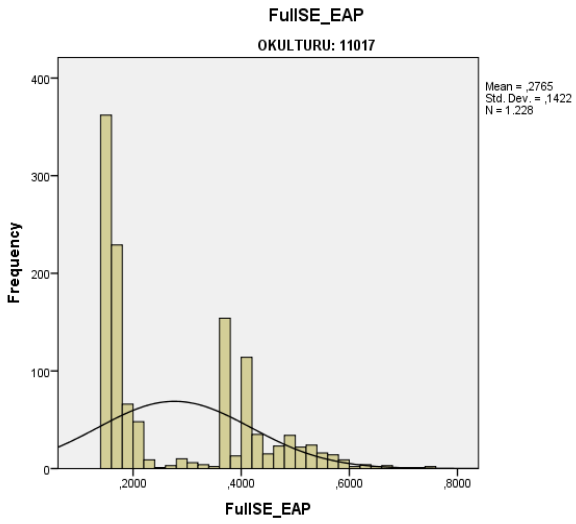


Figure 7. Distribution of standard errors by EAP P&P test for general high schools

Distribution of both CAT and P&P ability estimations by MLE obtained from Anatolian schools

Table 9 demonstrates the distributional features of Anatolian schools ability estimates from CAT simulations and P&P test by MLE. The findings showed that the median was higher than the mean by MLE P&P test for Anatolian schools. The findings of the simulation were stable in between different test termination conditions and CAT produced lower ability estimates than the conventional P&P test.

Table 9
Distribution of both CAT and P&P ability estimations by MLE obtained from Anatolian schools

		SE Thresholds		Fixed number of items		
	P&P	0.30	0.20	10	15	25
Mean	-.07	-.28	-.28	-.28	-.23	-.25
Median	.08	-.18	-.13	-.19	-.14	-.10
Mode	-3.45	.37	.17	.19	.16	-.00
Std. Deviation	1.08	.74	.67	.76	.73	.70
Skewness	-1.16	-.94	-1.10	-1.08	-1.05	-1.19
Kurtosis	2.50	.97	1.62	1.42	1.44	1.98

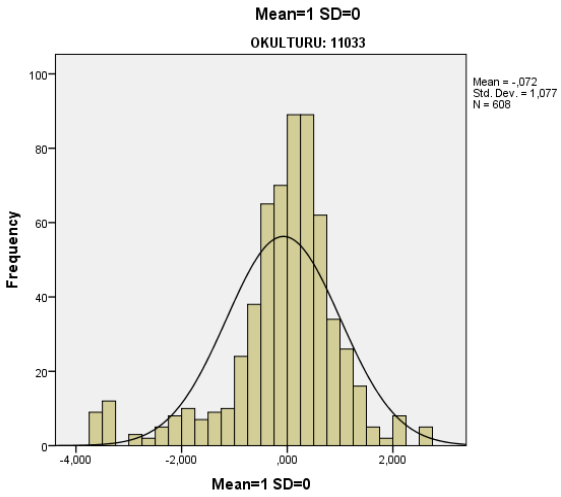


Figure 8.Distribution of MLE P&P ability estimates for Anatolian schools

Figure 8 represents the distribution of MLE P&P ability estimates for Anatolian schools. It can be clearly seen that most of the simulees had ability estimates around 1.00 to 2.00. This histogram had one peak point. The mean was - 0.07 whereas the median was 0.08 for Anatolian schools by MLE P&P test. The mean of the ability estimates was 0.23 and the median was 0.63 by MLE P&P test for general high schools. When the results obtained from general high schools were compared to results obtained from Anatolian schools, it can be said that the mean value of Anatolian schools' ability estimates were lower.

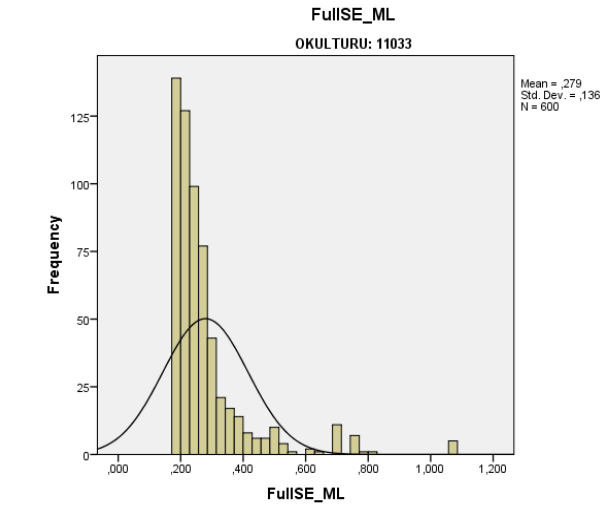


Figure 9. Distribution of standard errors by MLE P&P test for Anatolian schools

Figure 9 shows the distribution of SEs by MLE P&P test for Anatolian schools. The mean of the SE by MLE P&P test for Anatolian schools was 0.28. SE values were mostly between 0.20 and 0.40 which made the results highly reliable. As mentioned before, Anatolian schools have successful students. For this reason, SE results were very close to each other, mostly 0.20, when compared to general high schools. Clearly, SE 0.20 works well for simulees belonging to Anatolian schools. The histogram was again right-skewed since it had a tail on the left. Most of the simulees

had SE scores less than 0.40 so the simulation created more reliable results for Anatolian schools than for general high schools.

Distribution of both CAT and P&P ability estimations by EAP obtained from Anatolian schools

Table 10 shows that the median of ability estimations of the simulation were lower than the P&P test but the results obtained from simulation were similar to each other. The mean of the ability estimations was -0.03 and the median of the ability estimations were 0.47 for Anatolian schools by EAP P&P test. On the other hand, the mean of the ability estimations was 0.22 and the median of the ability estimations were 0.69 for general high schools by EAP P&P test. As seen, EAP produced lower ability estimations for Anatolian schools than it produced for general high schools.

Table 10
Distribution of both CAT and P&P ability estimates by EAP obtained from Anatolian schools

	P&P	SE Thresholds		Fixed number of items		
		0.30	0.20	10	15	25
Mean	-.03	.01	.01	.00	.00	.01
Median	.47	.35	.40	.22	.25	.34
Mode	-3.25	-3.01	-3.01	.47	-2.70	-2.91
Std. Deviation	1.02	1.01	1.03	.96	.98	1.01
Skewness	-1.56	-1.11	-1.12	-.91	-1.02	-1.12
Kurtosis	2.56	1.64	1.50	.91	1.26	1.44

The histogram given in Figure 10 demonstrates the distribution of EAP P&P ability estimations for Anatolian schools. Figure 10 was slightly left-skewed since there was a tail on the right. Most of the simulees were seemed between ability estimations of 0.70 to 1.50 so most of them were above the mean value of P&P test.

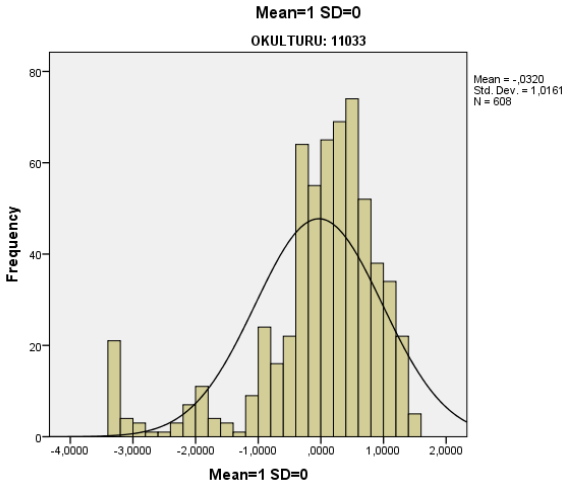


Figure 10. Distribution of EAP P&P ability estimates for Anatolian schools

Figure 11 shows the distribution of SEs by EAP P&P test for Anatolian schools. The mean of SE was 0.24. As can be seen from the Figure, most of the simulees had SE values less than 0.30 which made the test perfectly reliable. The histogram indicated that SE results were between 0.10 and 0.30 for EAP P&P test. Hence, most of the results were around SE 0.30. There were only a few simulees who had SE results more than 0.30 up to 0.50. For general high schools, there were some results up to SE with 0.80 by EAP but it can be aptly said that SE with 0.30 by EAP works perfect for Anatolian schools.

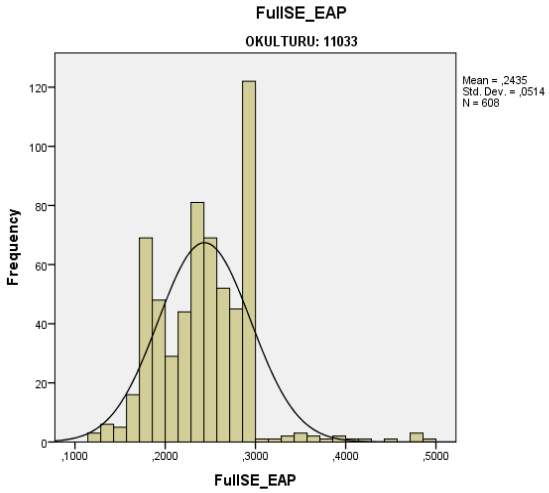


Figure 11. Distribution of standard errors by EAP P&P test for Anatolian schools

Distribution of both CAT and P&P ability estimations by MLE obtained from private schools

Table 11 demonstrated that MLE ability estimates had a distribution which was close to normal. The results were stable but CAT produced slightly lower medians than the conventional P&P test under different test termination rules. The mean and median values for MLE P&P test ability estimates of private schools were higher than both general high schools' and Anatolian schools'. Moreover, CAT produced lower ability estimations than the conventional P&P test for private schools by MLE. Also simulation results were similar to each other.

Table 11
Distribution of both CAT and P&P ability estimates by MLE obtained from private schools

	P&P	SE Thresholds		Fixed number of items		
		0.30	0.20	10	15	25
Mean	.01	.01	-.00	.06	.04	.01
Median	.13	.04	.08	.10	.11	.10
Mode	-2.58	-.10	-3.00[b]	.01[b]	-.93[b]	-3.00
Std. Deviation	1.02	.81	.83	.77	.79	.82
Skewness	-.67	-.95	-1.01	-.37	-.50	-.93
Kurtosis	.44	2.37	2.06	-.33	.39	2.07

b. Multiple modes exist. The smallest value is shown.

Figure 12 showed that private schools had a distribution of ability estimates by MLE P&P test that were close to symmetrical. The mean of the ability estimations were 0.01. Most of the simulees had ability estimates between 0.30 and 1.00. The range was between -3.00 and +3.00 as expected. The mean by MLE P&P ability estimations for Anatolian schools were - 0.07. On the other hand the mean of MLE P&P ability estimates for general high schools was 0.23, which led to conclude that the mean of private schools' ability estimations were between general high schools and Anatolian schools.

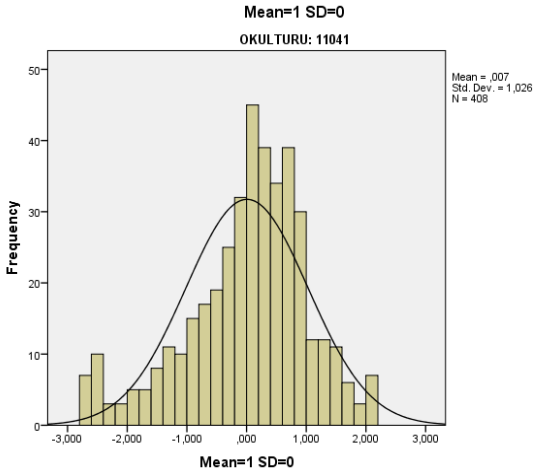


Figure 12. Distribution of MLE P&P ability estimates for private schools

Figure 13 demonstrates the distribution of SEs by MLE P&P test for private schools. The mean of SE was 0.27. The histogram was right-skewed hence the results were between 0.00 and 0.50. SE findings by MLE P&P test for private schools were not reliable enough since a handful of them were above SE 0.40. The results obtained from CAT simulations showed that MLE P&P test did not work properly for private schools. As it was up to SE with 2.50 for general high schools and it was up to SE with 0.80 for Anatolian schools, MLE may not be preferable for these schools types.

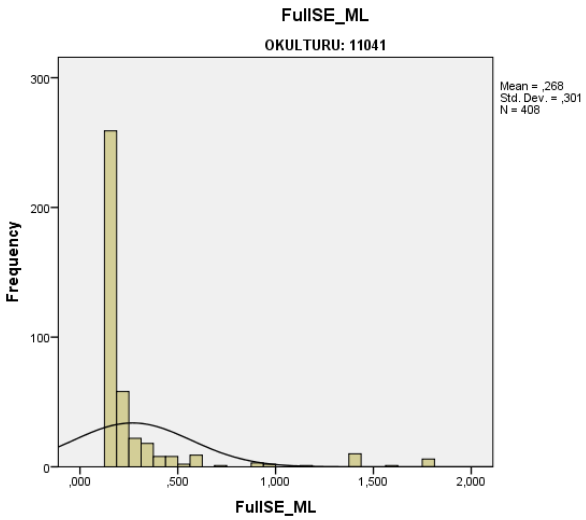


Figure 13.Distribution of standard errors by MLE P&P test for private schools

Distribution of both CAT and P&P ability estimates by EAP obtained from private schools

Table 12 indicated that, in general private schools produced lower ability estimates by simulation than conventional P&P test by EAP. The median value was higher than the mean but the results were stable in between different termination rules. For Anatolian schools, the mean and median of the ability estimations were -.03 and 0.47. As seen, these values were very close to private schools'. Since Anatolian and private schools had higher total scores from SSE 2005, they had more successful students than the general high schools'. For this reason, their results could be similar to each other. Also, SE 0.30 by EAP produced the closest result to the P&P test.

Table 12
Distribution of both CAT and P&P ability estimates by EAP obtained from private schools

	P&P	SE Thresholds		Fixed number of items		
		0.30	0.20	10	15	25
Mean	-.01	.07	.08	.08	.04	.06
Median	.38	.36	.40	.26	.24	.34
Mode	-2.34	-2.58	-2.58	-2.04	-2.24	-2.44
Std. Deviation	1.03	1.23	1.26	1.04	1.08	1.17
Skewness	-.72	-.34	-.35	-.26	-.33	-.36
Kurtosis	-.12	-.44	-.58	-.60	-.52	-.49

Figure 14 shows the distribution of EAP P&P ability estimates for private schools. The mean of ability estimations was -.01 for private schools whereas it was 0.02 for general high schools. Also the median of ability estimations was 0.07 for general high schools whereas it was 0.38 for private schools. As seen, ability estimates for private schools were lower than the general high schools' by EAP. Last, the histogram was left-skewed for EAP P&P ability estimates of Anatolian schools whereas it was right-skewed for EAP P&P ability estimates of general high schools.

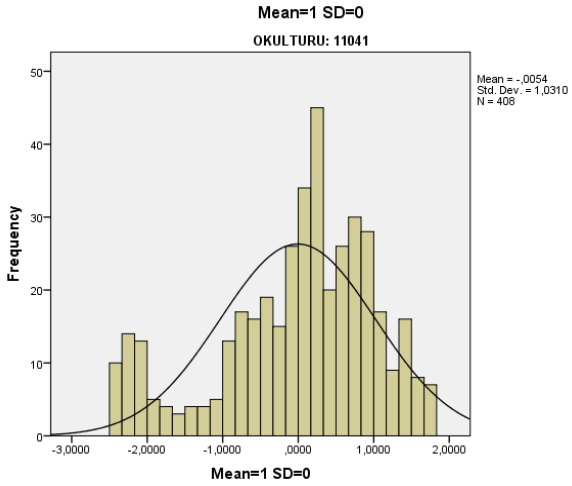


Figure 14. Distribution of EAP P&P ability estimates for private schools

Figure 15 represents the distribution of SEs by EAP P&P test for private schools. The mean of SE was 0.21 by EAP. Most of the simulees had SE scores between 0.10 and 0.30 which represents a high reliability. On the other hand, a handful of simulees had SE scores more than 0.40. The mean of SE was 0.24 by EAP P&P test for Anatolian schools whereas the mean value of SE was 0.28 and by EAP P&P test for general high schools. It seemed EAP produced SE values which were close to each other for all school types.

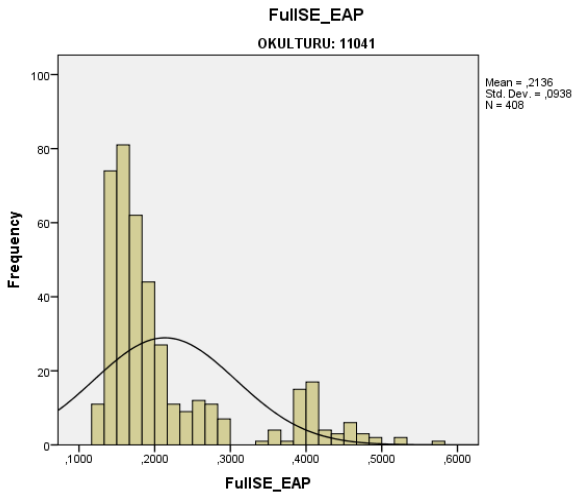


Figure 15. Distribution of standard errors by EAP P&P test for private schools

As it is assumed that three different school types have different ability levels, ability estimate methods may also differ. To be able to decide which one works better, it is important to check SE values for fixed items under MLE and EAP ability estimation methods.

Are there any differences in terms of the reliability of scores obtained from CAT and P&P test?

Table 13 shows post-hoc simulation results under fixed item termination rule by MLE and EAP ability estimation methods. As can be seen from the table, MLE ability estimates had somehow lower SE values than EAP estimates under fixed item termination rule. Hence, SE values of the fixed length tests decreased when the numbers of items increased for both MLE and EAP estimates.

Table 13
Median of SE values obtained from P&P and CAT under fixed items

School Type	P&P	MLE			EAP		
		10	15	25	10	15	25
General	0.24	.57	.13	.12	.43	.37	.25
Anatolian	0.24	.26	.23	.19	.39	.34	.29
Private	0.16	.28	.23	.19	.44	.41	.35

A closer look at the data indicated that the P&P test produced lower SE values than CAT, indicating the P&P version produced more reliable ability estimations. Table 13 shows that the median of SE value for 10 items was above 0.30 for general high schools by MLE. In addition, it was the maximum SE value seen in the Table 13. 15 and 25 fixed items for general high schools under MLE seemed as the minimum standard error value in the table. General high schools did not show stable results by MLE. Similarly 10 and 15 items did not produce SE values below 0.30 for EAP-based simulations. Clearly 15 and 25 fixed items simulations worked well for MLE.

Moreover, 25 fixed items produced SE values below 0.30 under EAP-based simulations, except for private schools. Fixed length tests showed that the one with 25 items by MLE for general high schools had the smallest SE, and the one with 10 items by MLE for general high schools had the largest SE value. For this reason, MLE may not be preferred for further applications since there is inconsistency between the results obtained from simulations.

For what percentage of test-takers is MLE not able to produce scores?

Although promising results were obtained from MLE, its usage may constitute a problem. Table 14 presents the percentage of examinees with non-converging ability estimates under different testing conditions.

Table 14
Percentages of examinees with non-converging ability estimates based on MLE

School Type	Test Termination Rules				
	Fixed SE		Fixed item test		
	0.30	0.20	10	15	25
General	33.7	33.5	34.1	34.0	34.6
Anatolian	23.0	23.0	24.2	23.9	23.7
Private	21.8	21.1	19.9	19.9	21.1

On the basis of the data currently available, non-converging seemed to be an important issue for both test takers and administrators. Table 14 demonstrated that general high schools had the highest rate of non-estimated examinees which was more than 34% due to the working principle of MLE. This situation may not be tolerated by the ÖSYM because 34 % is a huge number that means nearly half of the general high school students may not administer this adaptive test. On the other hand, other school types had also higher non-estimated examinees by MLE. For instance, Anatolian schools had a non-estimated examinee percentage of 24.2% whereas private schools had 21.8% . All values are too high to be accepted by the ÖSYM.

In contrast to MLE, EAP has no missing scores thanks to its working principle. MLE is disadvantaged in this situation. EAP works well in all tested conditions whether all the questions in a test are answered correctly or left blank. For this reason, diverging ability estimation can be handled by EAP. Although EAP looks favorable in all conditions, in order to improve CAT strategies, it would be better to examine other ability estimation methods.

Summary

In this chapter, analysis and results were explained in accordance with research questions in six sections. The first section (p.33) provided detailed information about the number of items given to examinees in post-hoc simulation under different CAT algorithms. In the second section (p.34), correlations between ability estimations from P&P and CAT simulations were explained. Then in the third section (p.35), the difficulty of CAT and P&P formats of the SSE was discussed. In the fourth section (p.36), different ability estimations obtained from different CAT simulations were categorized depending on the school types and ability estimation methods used. For each school type first MLE then EAP was analyzed and the results were presented. In the fifth section (p.49), the median of SE values for fixed items were investigated and details were provided. Lastly, in the sixth section (p.50), the issue of non-converging was presented and the percentages of examinees with non-converging ability estimates based on MLE were explained.

The results and interpretations are discussed in Chapter 5 with further implications and practices.

CHAPTER 5: DISCUSSION

Introduction

Overview of the study

The main purpose of the study was to explore if CAT could be an alternative to SSEP&P test. To do that, this study offered one main and six sub-research questions as written below:

Could student's scores from P&P and CAT formats be compared?

1. Is there any reduction in the number of items required by CAT?
2. Is there a correlation between ability estimates obtained from CAT and P&P tests?
3. Is there any difference in difficulty between the CAT and P&P tests?
4. Are there any differences in terms of score distributions obtained from CAT and P&P tests?
5. Are there any differences in terms of the reliability of scores obtained from CAT and P&P tests?
6. For what percentage of test-takers is MLE not able to produce scores?

The findings of the study were compared to the linkings in the literature components in the section of major findings. In addition, implications for practice, implications for further research and limitations were provided at the end of the chapter.

Major findings

Is there any reduction in the number of items required by CAT?

The strongest point of CAT prominent in the literature is that it provides reduction in the number of items required (Davey 2011; Kaptan, 1993; Zickar et al., 1998). The findings from the research showed that the number of items decreased significantly in CAT. SE threshold-based post-hoc simulations were investigated to assess if there was any reduction in the number of items used by CAT, and it was found that all tested SE values provided reduction in the number of items except SE of 0.10. In fact, SE with 0.30 and 0.20 by simulation required fewer items than SE with 0.10 in all tested conditions.

Indeed, SE 0.10 did not show any reduction for number of items required since the items required for that SE value were 45 (total number of items in original SSE science subtest) for both ability estimation methods. Although SE 0.10 has the maximum reliability (0.99 in CTT), it is not logical to use it in CAT.As these results have showed in line with Gokce (2012), Kalender (2011) and Tonidental (2001), the number of items increased when SE decreased since the computer needed to ask more items to estimate ability levels of test takers.

Different ability estimation methods, MLE and EAP, generated a wide variety of results. On the basis of the evidence currently available, it seems that MLE requires fewer items than EAP but it may not be enough because of validity issues. Moreover, the findings of the study revealed that SE 0.30 by EAP provided reliable results hence SE 0.30 was equal to 0.91 in CTT. The current literature on test termination abounds with the examples of SE 0.30 (Babcock & Weiss, 2012; Gokce, 2012; Kalender, 2011; Wall &Waltz, 2004; Weiss & Kingsbury, 1984) whereas other

researchers have used SE of 0.32 or 0.38 (Babcock & Weiss, 2009; İseri, 2002). The findings provided support for the key argument of the study that CAT calculated reliability using fewer items. That was the strongest point of CAT over the conventional P&P tests.

Is there a correlation between ability estimates obtained from CAT and P&P tests?

The correlations between the results obtained from CAT and P&P test of SSE in general high schools was the lowest in MLE whereas private schools had the highest correlations for both ability estimation methods (MLE and EAP) and test termination rules. Of all the three school types, general high schools were the most represented among the test takers; 54% of test takers were from general high schools. Low level of ability, as evidenced by the median of total scores was very low, it can be said that general high schools mostly represent lower achievement levels. These students may be expected to show inconsistent test response pattern, which could be reasons for the lowest correlations of general high schools by MLE.

Private schools had the highest correlation for all ability estimates as they were higher achieving test takers. Private schools are paid schools, and families of the students are generally at high socio-economic status. Epple & Romano (1998) states that "in private schools, high-ability, low-income students receive tuition discounts, while low-ability, high-income students pay tuition premia" (p.33). For this reason, private schools have a larger range in terms of total scores. In the analyses, the median of total scores of SSE was higher for private schools than general high schools, which meant test items were more suitable for private school students than it was for general high school students.

Anatolian schools showed higher correlation than general high schools, close to private schools, because Anatolian school students are tested for acceptance. The median of total scores of SSE was higher for Anatolian schools than private schools. The evidence obtained from the SSE 2005 shows that Anatolian school students represent higher cognitive level than the other test takers. Thus CAT may be an efficient format for Anatolian schools, too.

The findings from the study illustrated parallel results to the literature regarding ability estimations, which show that EAP method has correlations larger than MLE. Correlations between estimates for EAP were invariant for all conditions under different school types and test terminations. Therefore, correlations between EAP estimates were higher than 0.90, which was very strong in terms of ability estimation, supported by Chen & Choi (2009) and Choi, Kim& Chen (2011). In a converse manner, correlations between estimates for MLE were showed large variations, across all conditions, which were in parallel to the results reported by Iseri (2002), Gokce (2012) and Kalender (2011).

Is there any difference in difficulty between the CAT and P&P tests?

The findings supported the model that CAT produced lower ability estimates than the P&P test. In other words, CAT provided difficult tests since CAT format gives only the tailored items to examinees, it is reasonable to expect that CAT would be more difficult. For instance, Anatolian schools had the lowest ability levels, produced by the simulations because Anatolian schools were provided items which met with their ability levels, which meant Anatolian high schools were given harder test than the other schools types. But the difference between ability estimates between CAT and P&P may be explained by grading approaches. P&P ability estimations are

calculated based on raw scores (ÖSYM, 2014). ÖSYM erases one correct answer for four false answers and has been using CTT analysis. On the other hand, CAT is grounded in IRT models so the method to calculate ability estimates is completely different.

Are there any differences in terms of score distributions obtained from CAT and P&P test?

It was seen that the mean values were lower than the median values for all school types under the conventional P&P test and different CAT simulations. Distributions of the scores were skewed for all school types in the histograms. The mean of SE obtained from general high schools by both MLE and EAP were above 0.40 which was too high to be considered reliable. Evidently CAT produced values which were invariant for all tested conditions. Also CAT produced lower ability estimations than the conventional P&P test by both MLE and EAP.

Are there any differences in terms of the reliability of scores obtained from CAT and P&P test?

Test termination strategies are important to provide efficiency in measurement (Weiss & Kingsbury, 1984). Termination criteria must be accurate to assess the ability estimations of the examinees. Three different reliability values were used as test termination rule to conduct this research. It was seen that SE values decreased when the numbers of items increased for both MLE and EAP ability estimation methods. For most of the simulees, the P&P test provided higher reliability. Here it is important to note that P&P test used 45 items to calculate SE values. On the other hand CAT produced SE values using fewer items than the P&P test.

In previous pages, it was stated that SE with0.10 was not working due to the requirement of having 45 items on the test. Although SE with 0.10 represents the highest reliability, it cannot be used. For this reason, SE values of 0.30 and 0.20 were used to conduct further analysis. For instance, results obtained from CAT for Anatolian schools showed higher variety in SEs under MLE. A handful of simulees had results higher than SE of 0.40. Although Anatolian schools had the highest true scores, they also had the highest SE values under MLE. This means that scores estimated for Anatolian high school students were not reliable by MLE.

The findings showed that EAP required more items than MLE for all SE thresholds whereas MLE used fewer items in all tested conditions. According to Thissen & Mislevy (2000), fixed length tests were easier to understand for all examinee groups and they were seemed to be fairer than variable-length tests. By this way, each examine have the chance to obtain a reliable ability estimate. It seems hard to obtain reliable ability estimates under fixed length test since there are different groups, which represent different ability levels. Also it is hard to interpret results above SE with 0.30 under the fixed length test termination rule. For these reasons, SE of 0.30 by EAP seems to be the optimum algorithm for the CAT applications of this data set.

For what percentage of test-takers is MLE not able to produce scores?

The results showed that MLE produced a high number of examinees whose ability levels were non-estimated, due to the condition dictated by MLE. To be able to start ability estimation, MLE requires 1 correct and 1 incorrect response. However for some of the students the CAT software did not find a response pattern to satisfy this condition. Hambleton & Swaminathan (1984) addressed this issue and underlined working principle of MLE. MLE did not work if the responses were provided full (all items correct) or blank (all items incorrect) by test takers. The results showed that

general high schools had the highest percentage of missing scores which was 34% by MLE. Due to this non-converging issue, MLE could not be preferable for CAT.

To summarize, this study introduced the concept of CAT as an alternative to SSE in Turkey. On the basis of currently available evidences of this study, it can be concluded that there is a significant reduction in the number of items required to produce an accurate score. Moreover, the findings indicated that there is comparability between CAT and P&P formats of SSE since the results showed higher correlations. Lastly, SE with 0.30 was suggested as a test termination rule together with EAP ability estimation method to provide optimum algorithm for CAT implementations.

Implications for practice

This study investigated (i) if CAT could be an alternative to SSE or not and (ii) how the scores obtained from CAT and P&P versions of SSE are comparable. For this purpose, a set of post hoc simulations were conducted by changing the method of ability estimations and test termination rules, using real students' responses. Different SEs and different text lengths were used as test termination rules. By using scores obtained from different administrations of CAT, the comparability of SSE with CAT was identified.

One of the disadvantages of P&P format of SSE is that it puts tremendous pressure on students since it is administered only once a year. CAT can be done many (three or four) times in a year to make people familiar with the system. Moreover, it helps to lessen students' anxiety about SSE. Participants' highest scores can be used for placement purposes. Then each subject area can be tested many times in different days so that examinees do not lose a whole year if they need to take the SSE again.

Also quick feedback can be provided right after the exam so that it does not take time to wait for the results of SSE. It is known that transportation of booklets and ensuring the security of them is important for SSE but this can be overcomed by CAT administration since it is a computer-based and adaptive test. Blind-guessing and cheating can also be prevented in this way. Lastly, CAT can provide interactive and innovative items which SSEP&P tests cannot provide. It may help to ask critical thinking questions more in a creative way.

In addition, Turkey has aimed to increase technology usage in classrooms through curriculum reforms. It is believed that technology would improve the success of education system and better conditions would be provided both in teaching and learning processes (Aksit, 2007). In like manner, technology can be used in assessments to provide better and friendly exams which better fits students' ability levels.

Implications for further research

In this study, items were used in dichotomous format which means there was only one correct response out of five. The other four were false. However, it would be better to use a polytomous format for practical applications. In a polytomous format, responses are categorized depending on the degree of correctness. Incorrect responses are given different weights and examinees who answer the same items but mark different incorrect responses may have different ability estimates. Based on these varieties, in the polytomous format, responses can be graded according to how close they are to the right answer. Also different algorithms can be designed. For instance, multidimensional patterns can be used instead of unidimensional patterns.

This lets the examinee change his/her answer, go back to the previous items and review those (Zheng et al., 2012).

The revealed study underlined EAP as favored over MLE for the tested conditions but it is important to highlight that there is a need for further investigations. In addition, the security of software programs, item bank, mode effect and the number of computers are important areas to be considered for further analysis.

In this study, SSE 2005 science items were used for simulations which are all multiple choice items. If CAT is used as an alternative to SSE in practice, clearly there is a need for larger item bank. Moreover, each item has to be classified based on its difficulty level. Otherwise, item exposure may happen which means examinees may be given the same items. In addition, if the item pool is not large enough, examines may be given the items that do not match with their ability levels.

Percentages of true, false and missing scores of SSE 2005 science items are given in Table 3 (p.28). The results obtained from this table show that the test items do not match with the ability levels of the test takers. However, it does not mean that the items are not well-qualified. This study did not focus on a detailed item analysis but it is important to note that the findings in Table 3 underline an important issue. A number of students gave incorrect answers and some of them left the responses blank, which suggests that sufficient learning may not have happened as a results of teaching. Teachers may need to implement curriculum more efficiently and school environments may need to be improved, especially in general high schools.

For the sake of discussion, it is important to scrutinize the issue of population. If SSE is changed into CAT application, firstly all stake-holders need to be informed. CAT has different administration procedures and algorithms. Depending on the test

termination strategies, test takers may be given different numbers of items or fixed length items in a limited time. It means some of the examinees may be given fewer items whereas some of them may be given more items. Test takers are also asked to answer items in front of the computers in centers. Consequently, test takers or their families may complain about this situation. For this reasons, explanation of CAT to the public is very important for practical applications. YDS (Foreign Language Testing) was done by computer for the first time in 2014. It was a classical computer-based test. Likewise, if more of the national exams are done by CAT, people become more familiar with CAT.

Limitations

Post-hoc simulations were used to identify comparability but simulations may not reflect the real situations. Although the scores obtained from SSE 2005 were used to conduct analysis, it was still a simulation which could provide limited information. In addition, to be able to develop CAT, large item bank is needed. Moreover, a well-qualified software program is also indispensable and it may cost large amount of money at the beginning.

It can be hard to suggest the optimum conditions for CAT administrations but the optimum condition can be decided depending on the purpose, item pool, and efficiency or test termination rules under different ability estimation methods.

Implications for educational decision makers and future researchers were explained in previous pages. Hopefully, the findings offer insights to future studies and CAT is investigated more to implement it efficiently.

REFERENCES

Akşit, N. (2007). Educational reform in Turkey. *International Journal of Educational Development, 27*(2), 129-137.

Ayberk, C. (2012). *A comparison of psychometric properties of a general ability test which administered in P&P and computer based form* (Unpublished Doctoral dissertation), Ankara University, Turkey.

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *The Journal of Computerized Adaptive Testing, 1*(1), 1-18.

Baykul, Y. (1979). *Örtük özellikler ve klasik test kuramları üzerine bir karşılaştırma* (Unpublished Doctoral dissertation), Hacettepe University, Turkey.

Beguin, A. A., & Glas, C. A. W. (2001).MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66* (1), 541-562.

Bennet, R. E. (1998). *Reinventing assessment: speculations on the future of large-scale educational testing, a policy information perspective.* Princeton: Educational Testing Service.

Berberoglu, G. (2012). Üniversiteye giriş nasıl olmalıdır? *Cito Türkiye, 12*(16). 15-18.

Berberoğlu, G. (1988). *Seçme amacıyla kullanılan testlerde Rasch modelin katkıları* (Unpublished Doctoral dissertation), Hacettepe University, Turkey.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37* (1). 29-51.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* (Unpublished Doctoral Dissertation), The University of Texas at Austin, USA.

Chen, J., & Choi, J. (2009). A comparison of maximum likelihood and expected a posteriori estimation for polychoric correlation using Montecarlo simulation. *Journal of Modern Applied Statistical Methods, 8*(1), 337-354.

Choi, J., Kim, S., & Chen, J. (2011). A comparison of maximum likelihood and Bayesian estimation for polychoric correlation using Montecarlo simulation. *Journal of Educational and Behavioral Statistics, 36*(4), 523-549.

Christensen, R. (2002). Effects of technology integration education on the attitudes of teachers and students. *Journal of Research on Technology in Education*, *34*(4), 411-433.

Clark, R.E. (1994). Media will never influence learning. *Educational Technology, Research and Development*, *42*(2), 21-29.

Çıkrıkçı-Demirtaşlı, N. (2003). A study of raven standard progressive matrices test's item measures under classic and item response models: an empirical comparison. *Journal of Faculty of Educational Sciences, 35*(1-2), 71-79.

Davey, T. (2011). *A guide to computer adaptive testing systems*. Washington DC: Educational Testing Service.

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

Drasgow, F., & Chuah, S. C. (2006). Computer-based testing. In Eid M, Diener E (Eds.) *Handbook of multimethod measurement in Psychology* (pp. 87–100). Washington, DC: American Psychological Association.

Eignor, D. R. (1993). *Deriving comparable scores for computer adaptive and P&P tests: An example using the SAT.* Princeton, NJ: Educational Testing Service.

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

Epple, D., & Romano, E. (1998). Competition between private and public schools, vouchers, and peer group effects. *American Economic Review*, *88* (1), 33-62.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement,58*(3).357-382.

Fraenkel, J., & Wallen, N. (2009).*How to design and evaluate research in education.* New York: McGraw-Hill Education.

Fetzer, M., Dainis, A., & Lambert, A. (2011). SHLPreVisor*: Computer adaptive testing (CAT) in an employment context.* Retrieved from: https://central.shl.com/SiteCollectionDocuments/-%202011.pdf

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Gökçe, S. (2012). *Comparison of linear and adaptive versions of the Turkish pupil monitoring system (PMS) mathematics assessment* (Unpublished Doctoral dissertation), Middle East Technical University, Turkey.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12(*3), 38-47.

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

Hambleton, R. K., Zaal, J.N., & Pieters, Jo. P. M. (1991). *Advances in educational and psychological testing: Theory, applications and standards.* Netherlands: Springer.

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

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: bilgisayar ortamında bireyselleştirilmiş testler (computerized adaptive tests-CAT). *Cito Eğitim: Kuram ve Uygulama, 5*, 39-48.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. Doctoral dissertation, Middle East Technical University, Turkey.

Kalender, İ. (2012). Computerized adaptive testing for student selection to higher education. *Yükseköğretim Dergisi, 2*(1), 13-19.

Kaptan, F. (1993). *Yetenek kestiriminde bireyselleştirilmiş test uygulaması ile geleneksel kağıt-kalem uygulamasının karşılaştırılması* (Unpublished Doctoral dissertation), Hacettepe University, Turkey.

Keller, L. A. (2000). *Ability estimation procedure in computerized adaptive testing*. Technical report: AICPA Research consortium.

Kim, D. H., & Huynh, H. (2007). Comparability of computer and paper-pencil versions of algebra and biology assessments. *Journal of Technology, Learning and Assessment, 6*(4).5-30.

Kingsbury, G., & Zara, R. A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-379.

Lee, J., Park, S., & Kim, K. (2012).Web-based adaptive testing system (wats) for classifying students academic ability. *Turkish Online Journal of Distance Education, 13*(4). 25-35.

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.

McBride, J. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*(1), 121-140.

Mead, A. D., & Meade, A. W. (2010). Item selection using CTT and IRT with unrepresentative samples. Paper presented at the twenty-fifth annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA.

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

Ministry of National Education (MoNE). 2015. *TEOG*. Retrieved from http://www.meb.gov.tr/teog-sinavi-icin-her-sey-hazir/haber/8594/tr

Organisation for Economic Co-operation and Development (OECD). 2012. *Country note: Turkey.* Retrieved from http://www.oecd.org/pisa/keyfindings/PISA-2012-results-turkey.pdf

ÖSYM (2005). *Yüksek ögretime giriste okul türü ve ögrenim durumuna gore okul türü ve ögrenim durumuna gore basvuran yerlesen aday sayıları.* Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

ÖSYM (2008). *Yüksek ögretime giriste okul türü ve ögrenim durumuna gore okul türü ve ögrenim durumuna gore basvuran yerlesen aday sayıları*. Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

ÖSYM (2009). *Yüksek ögretime giriste okul türü ve ögrenim durumuna gore okul türü ve ögrenim durumuna gore basvuran yerlesen aday sayıları*. Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

ÖSYM (2014a). *Kamu personeli seçme sınavı lisans sonuçları.* Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

ÖSYM (2014b). *2014 ÖSYS klavuzu.* Retrieved from http://dokuman.osym.gov.tr/pdfdokuman/2014/YGS/2014_OSYS_KILAVUZU_02_01_2014.pdf

ÖSYM (2014c). *Yüksek ögretime giriste okul türü ve ögrenim durumuna gore okul türü ve ögrenim durumuna gore basvuran yerlesen aday sayıları*. Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

ÖSYM (2015). *Yüksek ögretime giriste okul türü ve ögrenim durumuna gore okul türü ve ögrenim durumuna gore basvuran yerlesen aday sayıları.* Ögrenci Seçme ve Yerlestirme Merkezi, ÖSYM Yayınları: Ankara.

Özbaşı, D. (2014). Bilgisayar okuryazarlığı testinin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliğine ilişkin bir araştırma (Unpublished Doctoral dissertation), Ankara University, Turkey.

Pashley, P.J. (1991). *An alternative three-parameter logistic item response model*. Educational Testing Service, N.J: Princeton.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184.

Riley, B., & Carle, A. (2012). Comparison of two Bayesian methods to detect mode effects between paper-based and computerized adaptive assessments: a preliminary Monte Carlo study. *BMC Medical Research Methodology*, *12*(124). 21-42.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software. 17*(5).1-25.

Rudner, L. (2012). *Demystifying the GMAT: Computer-based testing terms*. Reston: Graduate Management News.

Samejima,F. (1969). Estimation of latent ability using a response pattern graded scores. *Psychometrica Monograph*, 17.

Schaeffer, G., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993).*Field test of a computer-based GRE general test.* Princeton, NJ: Educational Testing Service.

Schaeffer, M., Steffen, M., & Durso, R. (1995). *The Introduction and comparability of the computer adaptive GRE general test*. New Jersey: Educational Testing Service.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*, 263-277.

Stocking, M., Smith, R., & Swanson, L. (2000*). An investigation of approaches to computerizing GRE subject tests.* New Jersey: Educational Testing Service.

Turkish Statistical Institute. (2014). *İstatistiklerle gençlik.* Retrieved from http://www.tuik.gov.tr/PreHaberBultenleri.do?id=16055

Thissen, D. & Mislevy, R. J. (2000). *Testing algorithms. In Wainer, H. (Ed). Computerized adaptive testing*. Mahwah: NH, Erlbaum.

Tindar, G., Haladyna, T.M. (2002). *Large-scale assessment programs for all students: validity, technical adequacy and implementation.* Mahwah: NH, Erlbaum.

Tonidental, S. (2001). Computer adaptive testing: The impact of test characteristics on perceived performance and test takers` reactions (Unpublished Doctoral dissertation), Rice University. USA.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2000).*Computerized adaptive testing: Theory and practice.* Norwell, MA: Kluwer.

Van der Linden, W. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp 3-30). New York: Springer.

Veldkamp, B. P., &Matteucci, M. (2013). Bayesian computerized adaptive testing.*aval. pol. públ. Educ., Rio de Janeiro, 21*(78), 57-82.

Verschoor, A. J. & Straetmans, G. J. J. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*, (pp 137-149). New York: Springer.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education*, *53*, 53-79.

Wainer, H. (1993). *Differential item functioning*. New Jersey: Erlbaum.

Wall, J.E., & Waltz, G. R. (2004).*Measuring up: Assessment issues for teachers, counselors, and administrators*. NC: National Board of Certified Counselors.

Wan, L., Keng, L., Davis, L., & McClarty, K. (2009). Methods of comparability studies for computerized and paper-based tests. *Test, Measurement & Research Service, 9* (10).1-4.

Wang, H. B., Kuo, B. C., Tsai, Y., & Liao, C. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *TOJET: The Turkish Online Journal of Educational Technology*, *11*(4). 1-13.

Wang, H., & Shin, D. D. (2010). Comparability of computerized adaptive and P&P tests. *Test, Measurement & Research Service,10*(13).1-7.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*, 5-24.

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

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*, 19–49.

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

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

Yi, Q., Wang, Y., & Ban, J. C. (2000). *Effects of scale transformation and test termination rule on the precision of ability estimates in CAT*. Iowa: ACT Report Series.

Yildirim, I., & Ergene, M. (2007). High rates of depressive symptoms among senior high school students preparing for national university entrance examination in Turkey. *The International Journal on School Disaffection, 1,* 35-43.

Yildirim, I. (2004). Test anxiety, daily hassles and social support as predictors of depression. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 27*(4). 241-250.

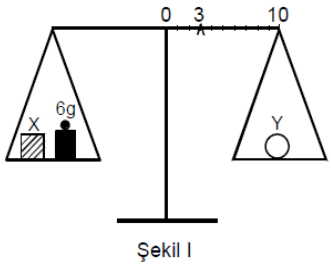
Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: design, heuristic assembly, and comparison with other testing modes, 12*(6). Iowa: ACT Research Report Series.

Zickar, M. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science, 7*(4), 104-109.

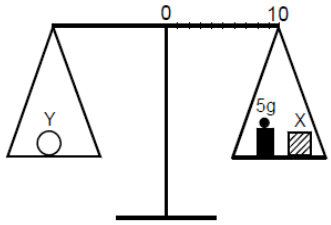
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago, IL: Scientific Software International.

APPENDIX: SSE 2005 Science items

46)



Şekil I

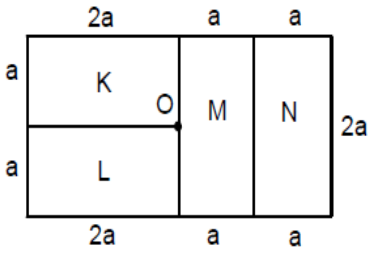


Şekil II

Eşit kollu bir terazinin kefelerinde Şekil I deki cisimler varken binici 3. Bölmeye getirilerek yatay denge sağlanıyor. Binicinin bir bölme yerdeğiştirmesi 0,1 g a denk geldiğine göre, terazinin kefelerinde Şekil II deki cisimler varken yatay dengenin sağlanması için binicinin kaçıncı bölmeye getirilmesi gerekir?

A) 4. B) 5. C) 6. D) 7. E) 8.

47)



Şekildeki levha, farklı metallerden yapılmış dikdörtgen biçimli, ince, düzgün ve türdeş K, L, M, N parçalarından oluşmuştur. Bu levhanın kütle merkezi O noktasıdır.

K, L, M, N parçalarının kütleleri sırasıyla m_K, m_L, m_M, m_N olduğuna göre,

I. $m_K = m_L$

II. $m_M = m_N$

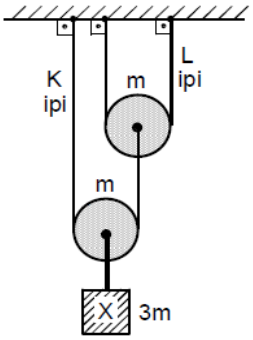
III. $m_K + m_L = m_M + m_N$

eşitliklerinden hangileri kesinlikle doğrudur?

A) Yalnız I B) Yalnız II C) I ve II

D) I ve III E) II ve III

48)

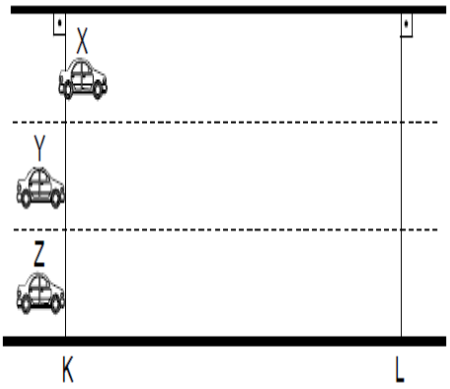


Şekildeki düzenekte X cisminin kütlesi 3m, makaraların her birinin kütlesi de m dir.

K, L iplerindeki gerilme kuvvetlerinin büyüklükleri sırasıyla T_K, T_L olduğuna göre, $\frac{T_K}{T_L}$ oranı kaçtır?

A) $\frac{1}{4}$ B) $\frac{1}{2}$ C) $\frac{2}{3}$ D) $\frac{3}{4}$ E) $\frac{4}{3}$

49)



Şekildeki doğrusal yolda X, Y, Z otomobilleri değişmeyen hızlarla KL yönünde gitmektedir. K çizgisinden önce X, sonra da Y ile Z aynı anda; L çizgisinden de önce Y, sonra da X ile Z aynı anda geçiyor.

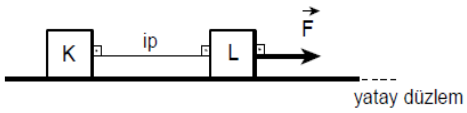
X, Y, Z otomobillerinin hızlarının büyüklükleri sırasıyla v_X, v_Y, v_Z olduğuna göre, bunlar arasındaki ilişki nedir?

A) $v_X = v_Z < v_Y$ B) $v_Y = v_Z < v_X$

C) $v_Z < v_X = v_Y$ D) $v_X < v_Y < v_Z$

E) $v_X < v_Z < v_Y$

50)



Şekildeki gibi iple birbirine bağlı K, L cisimleri sürtünmesiz yatay düzlemde, düzleme paralel sabit $\vec{F}$ kuvvetinin etkisinde hareket ederken ip kopuyor.

İp koptuktan sonraki süreçte, $\vec{F}$ kuvveti değişmediğine göre, K ve L nin hızlarının büyüklükleri için ne söylenebilir?

(Havanın etkisi önemsenmeyecektir.)

	K nin hızının büyüklüğü	L nin hızının büyüklüğü
A)	Azalır	Değişmez
B)	Azalır	Artar
C)	Değişmez	Değişmez
D)	Değişmez	Artar
E)	Artar	Artar

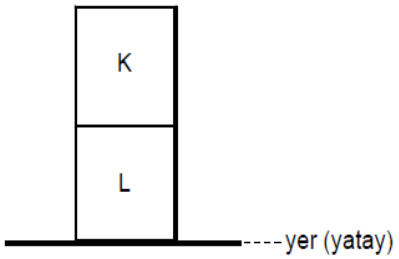
51)

Sabit $\vec{F}$ kuvveti, kütlesi 2 kg olan durgun bir cismi, düşey doğrultuda 15 m yükseltiyor ve bu cisme 10 m/s hız kazandırıyor.

Bu olayda, $\vec{F}$ kuvvetinin yaptığı iş kaç J dür?
($g = 10\ m/s^2$ olarak alınacak, havadaki sürtünme önemsenmeyecektir.)

A) 100 B) 200 C) 300 D) 400 E) 500

52)



Eşit hacimli türdeş K, L küpleri şekildeki gibi üst üste konulduğunda, yere göre potansiyel enerjileri birbirine eşit oluyor.

K nin özkütlesi d_K, L ninki de d_L olduğuna göre, $\frac{d_K}{d_L}$ oranı kaçtır?

A) $\frac{1}{4}$ B) $\frac{1}{3}$ C) $\frac{1}{2}$ D) $\frac{2}{3}$ E) 1

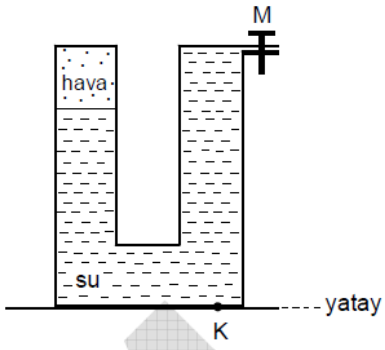
53)

Her birinin hacmi V olan K, L sıvılarının kütleleri sırasıyla m, 2m dir. Bu sıvıların tamamı karıştırılarak 2V hacimli türdeş karışım oluşturuluyor.

Karışımın özkütlesi d olduğuna göre, K sıvısının özkütlesi kaç d dir?

A) $\frac{1}{2}$ B) $\frac{2}{3}$ C) $\frac{3}{4}$ D) $\frac{4}{3}$ E) $\frac{3}{2}$

54)

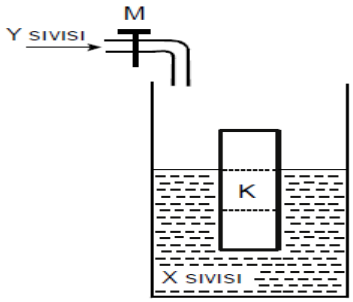


Düşey kesiti şekildeki gibi olan kabın M musluğu kapalıyken içindeki havanın basıncı P_h, K noktasında oluşan toplam basınç da P_K dir.

M musluğu açılınca dışarıya su aktığına göre, suyun aktığı süre içinde P_h ve P_K için ne söylenebilir?

	P_h	P_K
A)	Azalır	Azalır
B)	Azalır	Artar
C)	Artar	Azalır
D)	Artar	Değişmez
E)	Değişmez	Değişmez

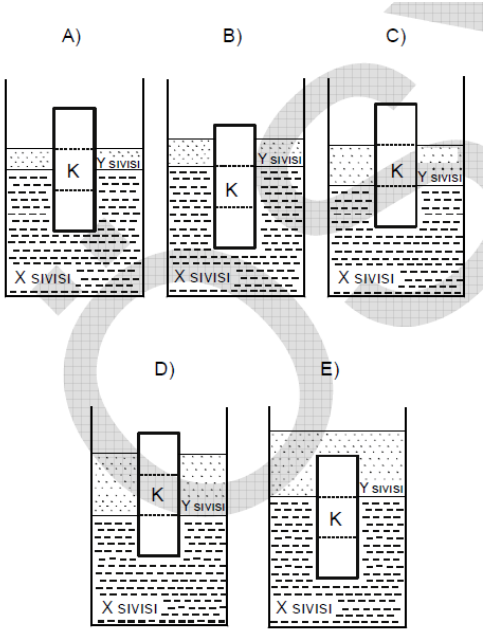
55)



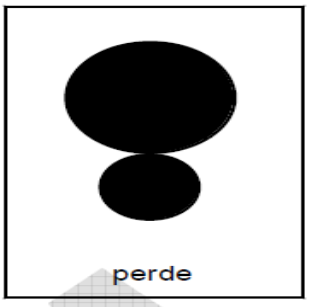
Eşit bölmeli K dik silindiri, bir kaptaki X sıvısı içinde şekildeki konumda dengede kalıyor. M musluğu açılarak, kaba özkütlesi X inkinden küçük olan Y sıvısı yavaş yavaş ekleniyor.

Y sıvısının eklenme süreci içinde, K silindirinin sıvılar içindeki görünümü aşağıdakilerden hangisi gibi olabilir?

(X, Y sıvıları karışmıyor. Sıcaklık değişimi yoktur.)



56)



Noktasal bir ışık kaynağı ile iki top bir perde önüne yerleştirilmiştir.

Perdedeki gölge şekildeki gibi olduğuna göre,

I. Topların yarıçapları birbirine eşittir.

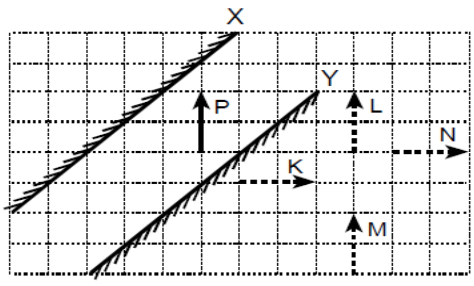
II. Topların merkezleri ışık kaynağına eşit uzaklıktadır.

III. Topların merkezleri ile ışık kaynağı aynı doğru üzerindedir.

yargılarından hangileri kesinlikle yanlıştır?

A) Yalnız I B) Yalnız II C) Yalnız III D) I ve II E) II ve III

57)

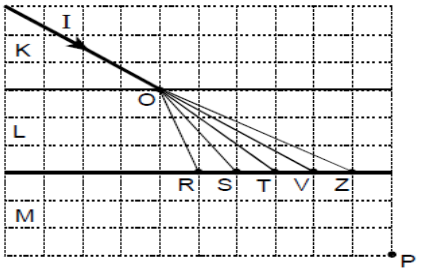


Birbirine paralel X ve Y düzlem aynaları arasına bir P cismi şekildeki gibi konuluyor.

Şekilde K, L, M, N ile belirtilenlerden hangi 2 si, P cisminin Y aynasındaki görüntüsüdür?

A) M ve N B) K ve L C) K ve M D) L ve M E) L ve N

58)

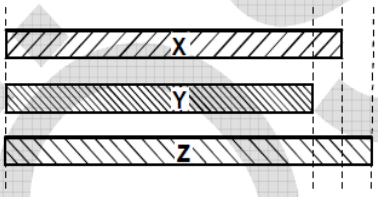


I ışık ışını, düşey kesitleri şekildeki gibi olan K, L, M saydam ortamlarından geçerek P noktasına ulaşıyor.

K ortamının ışığı kırma indisi M ninkine eşit olduğuna göre, bu ışının L ortamında izlediği yol aşağıdakilerden hangisidir?

A) OR B) OS C) OT D) OV E) OZ

59)



Uzama katsayıları birbirinden farklı olan X, Y, Z metal çubuklarının, T sıcaklığındaki boyları birbirine eşittir.

Bu çubuklara aşağıdaki işlemlerden hangisi uygulanırsa, çubukların görünümü şekildeki gibi olabilir?

A) Z yi T sıcaklığında tutarken, X i ve Y yi soğutma

B) Z yi T sıcaklığında tutarken, X i ısıtma, Y yi soğutma

C) Y yi T sıcaklığında tutarken, X i soğutma, Z yi ısıtma

D) X i T sıcaklığında tutarken, Y yi ve Z yi soğutma

E) X i T sıcaklığında tutarken, Y yi ve Z yi ısıtma

60)

Isıca yalıtılmış kapalı bir kaba, sıcaklıkları farklı, katı haldeki K, L maddeleri birbirine dokunacak biçimde konuluyor. Başlangıçta erime sıcaklığında olan L nin, ısıl denge kurulduktan sonra tümüyle eridiği gözleniyor.

Bu süreç sonunda

I. K nin sıcaklığı artmış, L ninki değişmemiştir.

II. K nin sıcaklığı azalmış, L ninki değişmemiştir.

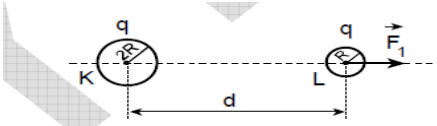
III. K nin sıcaklığı azalmış, L ninki artmıştır.

yargılarından hangileri doğru olabilir?

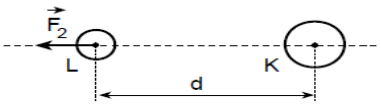
(Kaptaki havanın kütlesi önemsenmeyecektir.)

A) Yalnız I B) Yalnız II C) I ya da II D) I ya da III E) II ya da III

61)



Şekil I



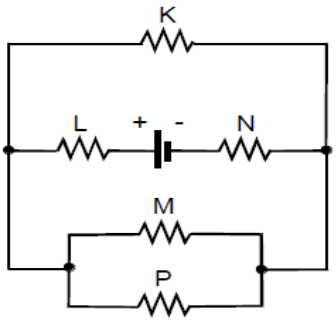
Şekil II

Şekil I deki iletken K, L kürelerinin yarıçapları sırasıyla 2R, R; elektrik yüklerinin büyüklüğü de q dur. Küreler Şekil I deki konumda tutulurken, L ye uygulanan elektriksel kuvvet $\vec{F}_1$ dir. Küreler birbirine dokundurulduktan sonra Şekil II deki konuma getirildiğinde ise L ye uygulanan elektriksel kuvvet $\vec{F}_2$ oluyor.

Buna göre, bu kuvvetlerin büyüklüklerinin $\frac{F_1}{F_2}$ oranı kaçtır?

A) $\frac{1}{2}$ B) 1 C) $\frac{9}{8}$ D) $\frac{4}{3}$ E) $\frac{3}{2}$

62)



Şekildeki elektrik devresi özdeş K, L, M, N, P dirençlerinden oluşmuştur. Bu devrede K, L, M dirençlerinden sırasıyla i_K, i_L, i_M şiddetinde elektrik akımları geçiyor.

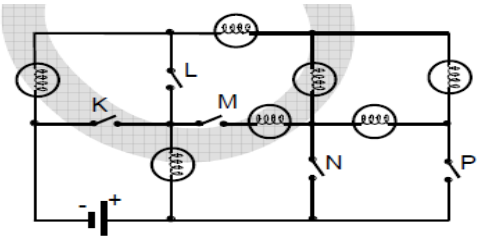
Buna göre, i_K, i_L, i_M arasındaki ilişki nedir?
(Üretecin iç direnci önemsenmeyecektir.)

A) $i_K = i_M < i_L$ B) $i_K = i_L < i_M$

C) $i_L < i_K < i_M$ D) $i_L < i_K = i_M$

E) $i_M < i_K = i_L$

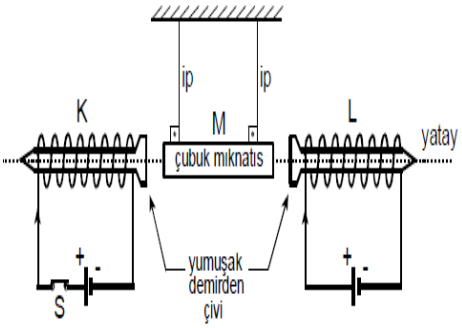
63)



Özdeş lambalardan oluşan şekildeki devrede açık olan K, L, M, N, P anahtarlarından hangisi kapatılırsa lambaların tümü ışık verebilir?

A) K B) L C) M D) N E) P

64)



İplerle asılı M çubuk mıknatısı, hareketsiz tutulan K, L elektromıknatıslarının etkisinde, şekildeki konumda dengede kalıyor.

Buna göre, S anahtarı açılarak K den geçen akım kesildiği anda M çubuk mıknatısı

I. Hareket etmez.
II. K ye doğru harekete başlar.
III. L ye doğru harekete başlar.

yargılarından hangileri doğru olabilir?

A) Yalnız I B) Yalnız II C) I ya da II
D) I ya da III E) II ya da III

65)

Aşağıdaki olaylardan hangisi molekül ya da atomların hareketiyle açıklanamaz?

A) Benzin dolu bidonun kapağı açılınca, benzin kokusunun odanın her tarafına yayılması

B) Bardaktaki suya damlatılan mürekkebin dağılarak suya renk vermesi

C) Bacalardan çıkan gazların havaya yayılması

D) Şişe mantarının suyun yüzeyinde kalması

E) Rüzgârlı havalarda rüzgâr gülünün dönmesi

66)

Bir madde, aşağıdaki özelliklerden hangisine sahipse arı madde değildir?

A) Belirli bir molekül formülünün olması

B) Tek cins atomlardan oluşması

C) Aynı cins atomlardan oluşan tek cins moleküllerden meydana gelmesi

D) Farklı cins atomlardan oluşan tek cins moleküllerden meydana gelmesi

E) Farklı cins moleküllerden, moleküller özelliklerini kaybetmeden ve aralarında belirli bir oran olmadan oluşması

67)

Aşağıdaki deneylerden hangisinin sonucunda gözlenen değişim, kesinlikle, karşısında belirtilen türden değildir?

	Deney	Değişimin türü
A)	Bir çözelti soğutulduğunda, içinde çözünmüş olan katının kristallenmesi	Kimyasal
B)	İki farklı arı sıvı oda koşullarında karıştırıldığında iki ayrı faz oluşması	Fiziksel
C)	İki farklı iyonik katının sulu çözeltileri karıştırıldığında çökelme oluşması	Kimyasal
D)	İki farklı sıvı karıştırıldığında gaz çıkışı olması	Kimyasal
E)	Bir katı madde ısıtıldığında gaz çıkışı olması	Kimyasal

68)

Cu ve Ag metallerinden oluşan bir alaşımdan alınan bir miktar örnek 0,1 mol Cu içermektedir. Bu örnek kapalı bir kapta yeterli miktarda H_2SO_4 ile tepkimeye girdiğinde kapta 0,2 mol SO_2 oluşmaktadır.

Cu ve Ag nin H_2SO_4 ile tepkimelerinin denkleştirilmiş denklemleri,

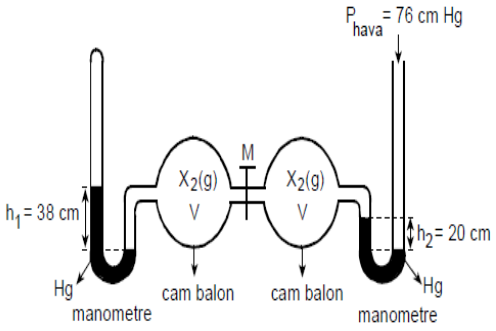
$$2Ag + 2H_2SO_4 \longrightarrow Ag_2SO_4 + 2H_2O + SO_2$$

$$Cu + 2H_2SO_4 \longrightarrow CuSO_4 + 2H_2O + SO_2$$

olduğuna göre alınan örnekteki Ag nin mol sayısı kaçtır?

A) 0,05 B) 0,10 C) 0,20 D) 0,25 E) 0,50

69)



X_2 gazıyla dolu özdeş cam balonlar birbirine ve manometrelere şekildeki gibi bağlanmıştır.

Sabit sıcaklıktaki bu sistemde M musluğu açıldıktan bir süre sonra, manometrelerdeki h_1 ve h_2 değerleri kaç cm olur?

	h_1	h_2
A)	29	58
B)	67	67
C)	32	47
D)	47	29
E)	58	29

70)

Bir X katısıyla hazırlanan ve aşağıda hacmi ile derişimi verilen doymamış sulu çözeltilerden hangisi, aynı koşullarda, <u>en az miktarda</u> X katısı ilavesiyle doymuş hale gelir?

	Çözelti hacmi (mL)	Çözelti derişimi (mol/L)
A)	5	1
B)	5	0,1
C)	5	0,5
D)	10	0,1
E)	10	1

71)

Aşağıdaki tabloda yapısı ve sudaki çözünürlüğü verilen maddelerden <u>eşit mol sayısında</u> alınmış ve alınan maddelerin her biri, <u>eşit hacimdeki</u> suyla ayrı birer kapta karıştırılmıştır.

Madde	Yapısı	Sudaki çözünürlüğü
İyot	Moleküler	Az
Üzüm şekeri	Moleküler	Çok
Gümüş klorür	İyonik	Çok az
Sodyum klorür	İyonik	Çok
Magnezyum klorür	İyonik	Çok

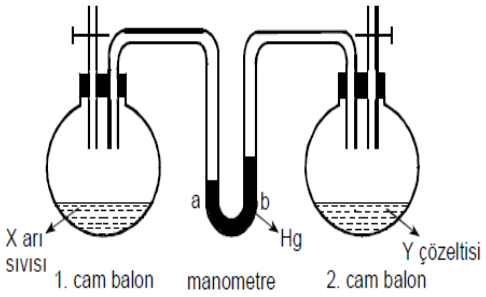
Bu maddelerin hangisiyle oluşturulan karışımın elektrik iletkenliği <u>en yüksektir</u>?

A) İyot B) Üzüm şekeri

C) Gümüş klorür D) Sodyum klorür

E) Magnezyum klorür

72)



Aynı sıcaklıkta içinde aynı hacimde sıvı bulunan özdeş 1. ve 2. cam balonlar, manometreye şekildeki gibi bağlanmıştır. 1. cam balonda X arı sıvısı, 2. cam balonda ise Y çözeltisi vardır. Y çözeltisi, uçucu olmayan bir katının X arı sıvısında çözünmesiyle oluşmuştur.

Bu sistemle ilgili,

I. Cam balonlar aynı anda özdeş ısıtıcılarla eşit ve kısa bir süre ısıtılırsa manometrenin a ve b kollarındaki cıva seviyeleri eşit olur.

II. 1. cam balona, X ile tepkime vermeyen kızgın bir metal parçası atılırsa manometrenin b kolunda cıva seviyesi yükselir.

III. 2. cam balona, aynı sıcaklıkta ve çözeltiyle tepkime vermeyen bir gaz eklenirse manometrenin a kolunda cıva seviyesi yükselir.

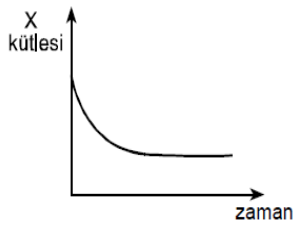
yargılarından hangileri doğrudur?

(I., II. ve III. işlemlerin birbirinden bağımsız olarak yapıldığı kabul edilecektir.)

A) Yalnız I B) Yalnız II C) Yalnız III

D) I ve II E) II ve III

73)



Ağzı açık bir kapta <u>yeterli süre</u> ısıtılan bir X maddesinin kütlesinin zamanla değişimi grafikteki gibidir.

Buna göre, X maddesi aşağıdakilerden hangisi olabilir?

A) Uçucu bir sıvının suyla oluşturduğu bir çözelti

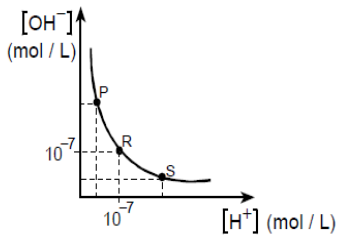
B) Havanın oksijeniyle birleşerek bileşik oluşturan bir metal

C) Birbiriyle tepkime vermeyen süblimleşen bir katıyla süblimleşmeyen iyonik bir katının karışımı

D) Birbiriyle tepkime vermeyen süblimleşen bir katının uçucu bir sıvıyla oluşturduğu bir çözelti

E) Isıtma ile tamamı iki farklı gaza dönüşen bir katı

74)



Sulu çözeltilerdeki OH^- derişiminin, H^+ derişimiyle ilişkisi grafikteki gibidir.

Bu grafiğe göre, sulu çözeltilerle ilgili aşağıdaki ifadelerden hangisi yanlıştır?

A) S noktasında Mg metali ile tepkimesinde H_2 gazı çıkar.

B) S noktasında kırmızı turnusol kâğıdını maviye çevirir.

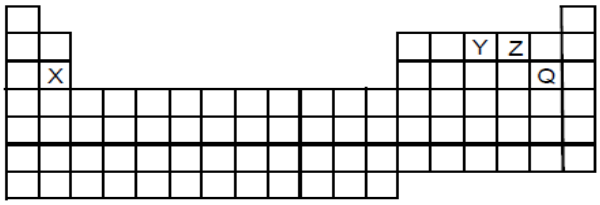
C) R noktasında nötrdür.

D) P noktasında HCl ile tepkime verir.

E) P noktasında elektrik akımını iletir.

75)

X, Y, Z, Q elementlerinin periyodik cetveldeki yerleri aşağıda gösterilmiştir.



Bu element atomları birbiriyle, aşağıda formülü verilen bileşiklerden hangisini oluşturur?

A) X_2Q_3 B) X_2Z_3 C) Y_2Q_3 D) Y_2Z_3 E) Z_2Q_3

76)

Aşağıda bazı element atomlarının son orbitalinin cinsi ve bu orbitaldeki elektron sayısıyla ilgili bilgiler ve bu elementlerle ilgili yargılar verilmiştir.

Verilen bilgilerden hangisinin karşısında belirtilen yargı yanlıştır?

	Element atomunun son orbitalinin cinsi ve bu orbitaldeki elektron sayısı	Elementle ilgili yargı
A)	d ; 6	Geçiş elementidir.
B)	p ; 2	Bileşiklerinde yalnız +2 değerlikli olur.
C)	p ; 5	Halojendir.
D)	2p ; 6	Atom numarası 10 dur.
E)	s ; 1	IA grubundadır.

77)

Tek atomlu olan X ve Y tanecikleri, aşağıdaki koşulların hangisinde birbirinin hem izotopu hem de iyonudur?

A) Yalnız proton sayıları eşit, nötron sayıları ve elektron sayıları farklı ise

B) Yalnız nötron sayıları eşit, proton sayıları ve elektron sayıları farklı ise

C) Yalnız elektron sayıları eşit, nötron sayıları ve proton sayıları farklı ise

D) Hem proton hem de nötron sayıları eşit, elektron sayıları farklı ise

E) Hem nötron hem de elektron sayıları eşit, proton sayıları farklı ise

78)

$$X + \alpha \longrightarrow Y + p$$

Yukarıdaki çekirdek tepkimesiyle ilgili aşağıdaki ifadelerden hangisi doğrudur?

A) X in proton sayısı Y ninkinden 2 fazladır.

B) Y nin nötron sayısı X inkinden 4 fazladır.

C) X ile Y nin molekül kütleleri eşittir.

D) Y nin atom numarası X inkinden 1 fazladır.

E) X ile Y nin kimyasal özellikleri aynıdır.

79)

. Aşağıdaki tabloda I, II, III, IV olarak numaralandırılan bakteri, mantar, bitki ve hayvan hücrelerinin bazı yapısal özellikleriyle ilgili bilgiler verilmiştir.

Hücresel yapılar / Hücreler	Kloroplast	Çekirdek zarı	Hücre duvarı ya da hücre çeperi
I	Yok	Var	Var
II	Var	Var	Var
III	Yok	Var	Yok
IV	Yok	Yok	Var

Buna göre, I, II, III, IV numaralı hücrelerin ait olduğu canlılar aşağıdakilerin hangisinde doğru olarak verilmiştir?

	Bakteri	Mantar	Bitki	Hayvan
A)	I	II	IV	III
B)	I	III	II	IV
C)	III	IV	I	II
D)	IV	I	II	III
E)	IV	II	III	I

80)

Hücrede gerçekleşen aşağıdaki olaylardan hangisi, enerji kullanılan bir metabolizma olayı değildir?

A) Karbondioksit difüzyonu

B) Glikozdan glikojenin oluşturulması

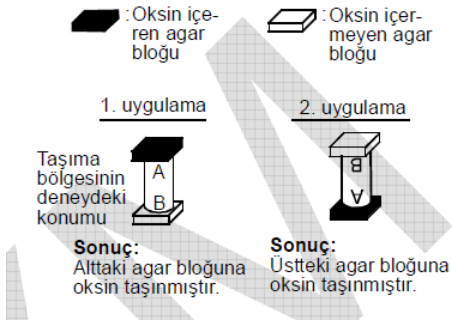
C) ADP nin ATP ye dönüştürülmesi

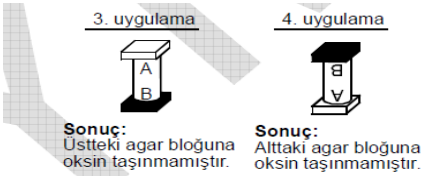
D) Klorofil taşıyan bir hücrede glikoz oluşturulması

E) Hücre zarında yıpranmış bölümlerin moleküler yapılarının yenilenmesi

81)

Bitkilerde tepe tomurcuğunda üretilen oksin (büyüme hormonu), bitkinin alt bölümlerine, tepe tomurcuğunun hemen altındaki taşıma bölgesiyle iletilir. Düzenlenen bir deneyde aynı bitkiden dört taşıma bölgesi kesilerek çıkarılmıştır. Deneydeki 1. ve 3. uygulamalarda kullanılan taşıma bölgeleri, bitkideki konumunda; 2. ve 4. uygulamalarda kullanılanlar ise ters konumda olacak şekilde, aşağıdaki şemada gösterildiği gibi, oksin içeren ve oksin içermeyen iki agar bloğu arasına yerleştirilmiş ve belirtilen sonuçlar alınmıştır.





Bu uygulamalardan elde edilen sonuçlara göre,

I. Taşıma bölgesinde, oksin hormonunun iletimi tek yönlüdür.

II. Yerçekimi kuvveti, oksin hormonunun taşınmasını sağlar.

III. Taşıma yönünü belirlemede taşıma bölgesindeki hücrelerin özelliklerinin rolü vardır.

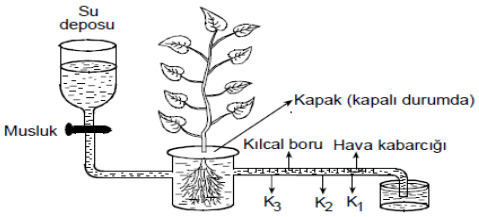
IV. Oksin hormonu bitkinin her bölgesine eşit olarak dağılır.

yargılarından hangilerine varılır?

A) I ve II B) I ve III C) I ve IV

D) II ve III E) III ve IV

82)



Bir bitkiyle şekildeki gibi bir deney düzeneği hazırlanmış ve düzeneğe su dolduktan sonra deponun musluğu kapatılmıştır. Deneyin başlangıcında kılcal borudaki hava kabarcığının bulunduğu K_1 noktası işaretlenmiştir. Deneye, karanlık bir ortamda başlanmış ve bir süre sonra ışıklandırılarak devam edilmiştir. Deneyde, hava kabarcığının ortamın karanlık olduğu süre sonunda K_2 noktasına kadar; ortamın ışıklandırılmasından sonraki süre sonunda ise K_3 noktasına kadar hareket ettiği görülmüştür.

Buna göre hava kabarcığının K_1 noktasından K_3 noktasına gelmesine neden olan su kaybı, bitkide gerçekleşen,

I. K_1 – K_2 arasında terleme,

II. K_1 – K_3 arasında solunum,

III. K_2 – K_3 arasında fotosentez

olaylarından hangileriyle açıklanır?

A) Yalnız I B) Yalnız II C) I ve III

D) II ve III E) I, II ve III

83)

Bir hayvan hücresinde, enzim sentezi sonucunda aşağıdaki moleküllerden hangisinin miktarı artar?

A) ATP B) tRNA C) Aminoasit

D) mRNA E) Su

84)

Normal olarak hücrelerde H_2O_2 , katalaz enzimiyle su ve oksijene parçalanır:

$$H_2O_2 \xrightarrow{\text{Katalaz}} H_2O + 1/2\ O_2$$

Bu olayla ilgili bir deneyde, karaciğer ve havuçtan alınan doku örneklerine aşağıdaki tabloda verilen işlemler uygulandıktan sonra bu örnekler, içinde eşit miktarda H_2O_2 bulunan 12 özdeş tüpe ayrı ayrı konulmuş ve tabloda belirtilen sıcaklıklarda tutulmuştur. Belirli bir süre boyunca tüplerdeki oksijen çıkışı gözlenmiş ve tabloda belirtilen bulgular elde edilmiştir.

	Karaciğer		Havuç	
Uygulanan işlem	Parça parça doğranmış	Ezilerek hücreleri parçalanmış	Parça parça doğranmış	Ezilerek hücreleri parçalanmış
Kaynatıldıktan sonra oda sıcaklığına getirilmiş doku + H_2O_2	O_2 çıkışı yok	O_2 çıkışı yok	O_2 çıkışı yok	O_2 çıkışı yok
Oda sıcaklığındaki doku + H_2O_2	O_2 çıkışı var	O_2 çıkışı var	O_2 çıkışı var	O_2 çıkışı var
0°C deki doku + H_2O_2	O_2 çıkışı yok	O_2 çıkışı yok	O_2 çıkışı yok	O_2 çıkışı yok

Bu deneyin bulgularına dayanarak,

I. Enzimin belirli sıcaklıklarda işlev görmesi için bozulmamış (kaynatılmamış) olması gerekir.

II. Enzim, hücre dışında da etkisini gösterir.

III. Enzimin bulunması olayın başlaması için yeterlidir.

IV. Enzim, belirli sıcaklıkların üstünde geri dönüşümü olmayan değişime uğrar.

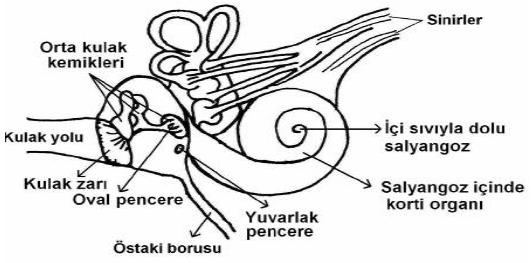
yargılarından hangilerine varılabilir?

A) I ve III B) II ve III C) I, II ve III

D) I, II ve IV E) II, III ve IV

85)

. Aşağıdaki şema, insan kulağında, ses dalgalarının beyne uyarı olarak iletimini sağlayan yapıları göstermektedir.



Aşağıdakilerin hangisinde, şemadaki yapılardan biri, gerçekleştirdiği işlevle birlikte verilmiştir?

A) Östaki borusu – Ses dalgalarının şiddetini artırma

B) Salyangoz – Orta kulak ile dış ortam arasında hava basıncını dengede tutma

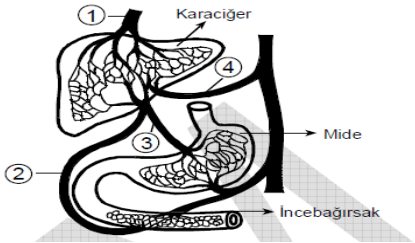
C) Kulak zarı – Havada yayılan ses dalgalarını sıvıda yayılan dalgalara çevirme

D) Orta kulak kemikleri – Ses dalgalarının şiddetinin aynı kalmasını sağlama

E) Korti organı – Farklı frekanslardaki ses dalgalarını impulsa çevirme

86)

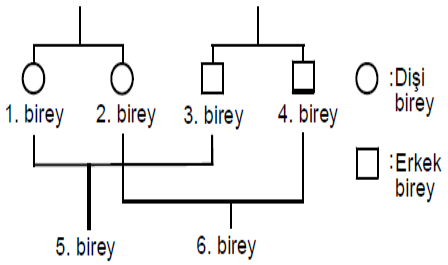
İnsanda, karaciğerin bazı besin maddelerinin depolanması, kanın zehirli maddelerden arındırılması, homeostasisin sağlanması gibi görevleri vardır. Aşağıdaki şemada, karaciğere kan getiren ve karaciğerden kan götüren damarlar numaralanarak gösterilmiştir.



Buna göre, karaciğere kan getiren ve karaciğerden kan götüren damarlar aşağıdakilerin hangisinde doğru olarak gruplanmıştır?

	Karaciğere kan getiren damarlar	Karaciğerden kan götüren damarlar
A)	1, 3	2, 4
B)	1, 4	2, 3
C)	1, 2, 3	4
D)	2, 3, 4	1
E)	4	1, 2, 3

87)



Yukarıdaki soyağacında, 1. ve 2. bireyler aynı yumurta ikizi, 3. ve 4. bireyler ayrı yumurta ikizidir.

Bu soyağacına göre,

I. 1. ve 2. bireylerin doku grupları aynıdır.

II. 3. ve 4. bireylerin kan grupları aynıdır.

III. 5. ve 6. bireylerin cinsiyetleri aynıdır.

IV. 1. bireydeki homozigot baskın özellikler 6. bireyin fenotipinde görülür.

yargılarından hangileri kesin olarak doğrudur?

A) I ve II B) I ve IV C) II ve III

D) I, III ve IV E) II, III ve IV

88)

Doğadaki azot döngüsünün bazı basamakları aşağıda verilmiştir:

I. Saprofit bakterilerin amonyak oluşturması

II. Denitrifikasyon bakterilerinin faaliyeti

III. Baklagil kök yumrucuklarındaki simbiyotik bakterilerin faaliyeti

Bu olayların hangi sırayla gerçekleşmesi, havadaki azotun canlı yapısına katılıp tekrar havaya dönmesini sağlar?

A) I – III – II B) II – I – III C) II – III – I

D) III – I – II E) III – II – I

89)

İndikatör (gösterge) tür, çevresindeki yararlı ya da zararlı maddelerden birine karşı çok duyarlı olan canlı türü olarak tanımlanır. Örneğin, kızböceklerinin bazı türleri, sudaki gelişim dönemlerinde, ortamdaki oksijenin azalmasına çok duyarlı olduğundan, bu böceklerin bulunduğu su ortamlarının temiz ve oksijen bakımından zengin olduğu söylenebilir.

Buna göre, bir türün indikatör (gösterge) tür olması için aşağıdaki özelliklerden hangisine sahip olması gerekir?

A) Ekolojik toleransının (hoşgörüsünün) az olması

B) Mutasyona uğrama sıklığının yüksek olması

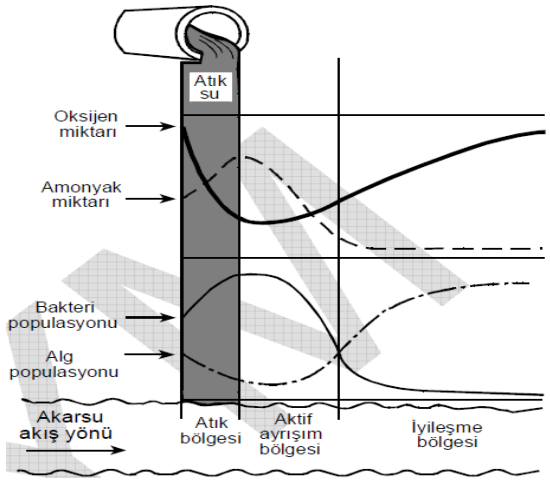
C) Hayat devresinin kısa olması

D) Metabolizma hızının yüksek olması

E) Populasyon büyüme hızının sınırlı olması

90)

0. Aşağıdaki grafik, atık su boşaltılan bir akarsu ortamında, atığın boşaltıldığı atık bölgesinden iyileşme bölgesine doğru gidildikçe, oksijen ve amonyak miktarları ile bakteri ve alg populasyonlarında meydana gelen değişiklikleri göstermektedir.



Yalnızca bu grafikteki bilgilere göre, bu akarsu ortamıyla ilgili olarak aşağıdakilerden hangisi söylenemez?

A) Oksijen miktarı ve bakteri populasyonu değişme eğrileri birbirine terstir.

B) Ortamda alglerin çoğalması, oksijen miktarındaki artışta rol oynar.

C) Bakteri ve alg populasyonları aynı besin maddelerini kullanır.

D) Ortama atık madde girmesi, alg populasyonunun azalmasına neden olur.

E) Amonyak miktarındaki değişmeler bakteri populasyonuyla ilgilidir.