Ampirik Araştırma / Empirical Research

www.yuksekogretim.org

Yükseköğretim Dergisi
Journal of Higher Education

Computerized Adaptive Testing for Student Selection to Higher Education*

Yükseköğretime öğrenci seçmek için bilgisayar ortamında bireyselleştirilmiş testler

İlker Kalender

Faculty of Education, Bilkent University, Ankara, Turkey

Özet

Bu çalışmanın amacı bilgisayar ortamında bireyselleştirilmiş test formatının Türkiye'deki yükseköğretime öğrenci seçme sınavlarına alternatif olarak kullanılabilirliğini tartışmaktır. Çalışmada, önce mevcut öğrenci seçme sistemindeki sorunlar ortaya konmaktadır. Mevcut sınav sistemi öğrencilerin üzerinde büyük bir baskı yaratmakta, ölçme-değerlendirme bakımından eleştiriye açık unsurlar içermekte ve sistemin kamuoyu tarafından eleştirilmesine neden olmaktadır. Bunun ardından bilgisayar ortamında bireyselleştirilmiş testlerin dayandığı temel prensipler ve sağladıkları avantajlar açıklanmakta ve Türkiye'deki öğrenci seçme sınavlarının bu formatta uygulanabilirliğini araştıran deneysel bulgular paylaşılmaktadır. Daha sonra ise, iki farklı araştırma deseni (simülasyon ve gerçek birey uygulaması) kullanan bir çalışmanın sonuçları paylaşılmaktadır. Sonuçlar (i) bilgisayar ortamında bireyselleştirilmiş formatın öğrenci seçme sınavının kağıt kalem formatı ile karşılaştırıldığında öğrencilere sorulan soru sayısında %80'e varan düşüşler sağladığını ve (ii) yetenek kestirimlerinin yüksek güvenilirliğe sahip olduğunu ortaya koymaktadır. Bireylerin klasik öğrenci seçme sınavlarından elde edilen yetenek kestirimleri ile CAT simulasyonlarından gelen kestirimleri arasındaki korelasyonlar 0.80'in üzerinde bulunmuştur. Gerçek bireyler ile yapılan CAT uygulaması da umut verici bulgular ortaya koymuştur. Çalışmanın sonunda bilgisayar ortamında bireyselleştirilmiş testlerin öğrenci seçme sistemi kullanımının mevcut sorunlara nasıl çözüm getirdiği tartışılmış, ayrıca bireyselleştirilmiş formata geçiş konusunda bir takım noktalara değinilmiştir.

Anahtar sözcükler: Bilgisayar ortamında bireyselleştirilmiş testler, öğrenci seçme ve yerleştirme sınavı, yükseköğretime giriş.

Abstract

The purpose of the present study is to discuss applicability of computerized adaptive testing format as an alternative for current student selection examinations to higher education in Turkey. In the study, first problems associated with current student selection system are given. These problems exerts pressure on students that results in test anxiety, produce measurement experiences that can be criticized, and lessen credibility of student selection system. Next, computerized adaptive test are introduced and advantages they provide are presented. Then results of a study that used two research designs (simulation and live testing) were presented. Results revealed that (i) computerized adaptive format provided a reduction up to 80% in the number of items given to students compared to paper and pencil format of student selection examination, (ii) ability estimations have high reliabilities. Correlations between ability estimations obtained from simulation and traditional format were higher than 0.80. At the end of the study solutions provided by computerized adaptive testing implementation to the current problems were discussed. Also some issues for application of CAT format for student selection examinations in Turkey are given.

Key words: Computerized adaptive testing, entrance to higher education, student selection and placement.

Large-scale testing is widely used in Turkey for many examinations such as the Student Selection and Placement Examination, the Foreign Language Examination for Civil Servants, and the Entrance Examination for Graduate Studies. These paper and pencil based tests are taken by thousands of individuals at the same

İletişim / Correspondence:
Dr. İlker Kalender
Bilkent University,
Faculty of Education,
Ankara, Turkey
Tel: +90 312 290 22 56
e-posta: kalenderi@bilkent.edu.tr

Yükseköğretim Dergisi 2012;2(1):13-19. © 2012 Deomed

Geliş tarihi / Received: Haziran / June 21, 2011; **Kabul tarihi / Accepted:** Mart / March 9, 2012;
Online yayın tarihi / Published online: Mart / March 30, 2012

*Part of the present study was presented at International Higher Education Congress 2011, May 27-29, Istanbul, Turkey.

Çevrimiçi erişim / Available online at: www.yuksekogretim.org • doi:10.2399/yod.12.004 • Karekod / QR code:

date using booklets including the same multiple-choice items. Among them the Student Selection and Placement Examinations (SSE) are of special importance since entrance to higher education programs mainly depends on these examinations. They are administered once a year with participation of over one million students and need a large organization including security of booklets, coordination of proctors and other test staff, transfer of test documents to test centers and get back to Student Selection and Placement Center, evaluation of answer sheets, etc.

Although there is a huge experience accumulated up to date on large scale testing in Turkey, there are problems that can be, and in fact were in the past, encountered.

Problems Associated with Current Student Selection Examinations (*not exhaustively*)

- Principle problem of the current system is the fact that one of turning points of students in their life is dependent on a multiple-choice test that is conducted once a year. That puts an extremely high pressure on students. Future life of students, whether they go on to higher education or not, is determined by an exam. Also if a student has a problem such as illness or he or she is late to exam due to some reasons out of control, there is nothing to do except waiting next year.
- Another major problem from measurement context is that item difficulty and ability levels of test-takers do not match. Therefore students may face items that are not appropriate for their levels in difficulty. Investigation of means of correct responses can provide information of that. For example, on first phase of the SSE in 2009, mean of science subtest is 4.0 out of 30 items (Student Selection and Placement Center, 2011). Those low mean scores indicate that students are not given proper items in terms of difficulty and there is problem of balancing difficulty and ability levels of test-takers.
- Unbalanced item difficulty shows itself in item and test parameters. A correct response results in a significant change in the students' ranks. A student who makes a blind guessing for a difficult item has a chance of 20% to give a correct response and by this way he or she can go up in ranking. Moreover students do not make blind guessing, rather they try to eliminate some of the alternatives before marking one. Therefore probability of giving correct response is higher than its expected value of 20%.
- Asking many items results in that students face very easy or hard items that are outside their ability levels. To cover a broad range of ability levels of examinees, many items are put exist in booklets. In other words, some of the items mean waste of time for students who take the test. Furthermore using more items may make students bored, fatigue, careless, etc.
- Steal or reveal of items in booklets constitutes a problem. This happened before; the booklets were revealed before the exam date and as a result exam was postponed. Security of booklets in whole country is not easy and there is potential of leakage of items. Postponement or cancellation of the exam might a good action to take, however, effect of that on students' psychological situations are not discussed. Also reveal of booklets results in a rise in public concern as to security issues. Limited public resources are wasted due to cost of postponement of the exams.
- Logistic of SSE is another problematic area. Test documents of students are collected and sent to Student Selection and Placement Center from all regions of Turkey. It is important to transfer test documents without any loss or damage. Despite all efforts, some sheets may be lost and damaged unintentionally. Although students are given a right to take the exams again without waiting next year, this solution puts another pressure on students.
- Since items are given in printed forms, there is a restriction in terms of item format. Only items that can be asked on paper can be used. That limits type of items that can be given. Items with multimedia components such as videos, animations with user interaction may help to achieve a better evaluation experience.

Problems listed above (i) exert pressure on students that results in test anxiety, (ii), produce measurement and evaluation experiences that can be seriously criticized, and (iii) lessen credibility of the student selection system and responsible institutions and make the student selection system questionable by public.

The present study seeks to investigate applicability of computerized adaptive testing (CAT) procedures as an alternative to SSE given as paper and pencil based. Advantages of CAT format are given from the perspective of problems of SSE in Turkey and also results of a study conducted experimentally related to applicability of CAT to SSE will be presented. Even tough the findings that will be presented in the study are related to SSE, they are expected to be valid for other large-scale test administrations given above.

Computerized Adaptive Testing (CAT)

The logic of computerized adaptive testing is based on idea of giving only appropriate items to individuals in difficulty from an item bank (Mead & Drasgow, 1993). Therefore individuals do not need to face items that they have a very low prob-

ability to give a correct answer or items that can be answered correctly with 100% chance.

Although adapting or tailoring tests for individuals is an old idea, rise of the practical application was not until 1980s. Weiss (1983) proposed computerized adaptive testing, stating that individuals' ability levels could be dynamically estimated during testing using computers. After each response given by an individual, computer can make an ability estimation using responses up to that point. After obtaining ability estimation, computer uses that ability for using next item from a large item bank. This process is repeated until test is terminated.

Details of a typical CAT session are as follows: First computer selects an item. Giving an item with moderate difficulty would be a good starting point or some a priori information such as students' high school scores can be used. Giving some items to obtain an initial estimate of ability is another option. After response to first item, depending on ability estimation method used for CAT, computer estimates initial ability or gives some more items prior to ability estimation. Once first ability estimation is obtained, computer starts to select more appropriate items for that test-taker. If a correct response is given to that item, computer updates ability estimation with a higher value and next item is selected from more difficult items appropriate for new ability estimate. If once again a correct response is given, computer decides that ability level of test-taker is higher and updates it. Increasing ability level goes on while correct responses are given. If test-taker gives a false response at a point in the test, computer decreases ability level and selects an easier item. By adapting the difficulty level of the test, computer tries to narrow range of ability level for test-taker. With each update in ability level, computer becomes surer about reliability of ability estimation. Test can be terminated when a predetermined reliability level is achieved. Test can also be ended after a certain number of items are given, on the other hand, that test termination rule sure does not assure an acceptable reliability level for ability estimation.

It is important to note that giving appropriate items does not mean each individual takes a test with moderate difficulty. Rather computer forces each test-taker by giving difficult items as long as correct responses are given. In a similar way, after wrong responses computer estimates a lower ability level and selects less difficulty items.

CAT format reduces number of items received by individuals by 50% compared to the paper and pencil format of the same test and also makes more reliable measurement experience possible (Embretson, 1996). This is the one of the major advantages of CAT administration.

In the literature, there are many advantages of CAT implementation stated by researchers (Cikrikci-Demirtasli, 1999; Embretson, 1996; Hambleton & Swaminathan, 1984; Rudner, 1998; Sands et al., 1997). Some of them are as follows: (i) Item difficulty of the test matches the ability level of individuals, and therefore test-takers do not encounter items very easy or very hard for them, and testing time shortens, (ii) There is no need to use printed test materials such as questions booklets, optical answer sheets, etc. Also process of transportation and using optical readers is eliminated, (iii) Since CAT is a dynamic process, scores of individuals are delivered immediately after the test is terminated, (iv) Item format is not restricted to paper-based questions. New item formats including, for example, multimedia or hotspots can be used.

A Study about CAT Implementation for Student Selection Examination

In that section, results of a study conducted by Kalender (2011) that investigated applicability of CAT for SSE science subtest are presented. In that study, science subtest of SSE was used. On the other hand, findings of that study can easily be generalized to other SSE subtests (mathematics, social sciences and Turkish) since profile of test-takers, item formats, etc. are similar across subtests.

Implementation of CAT for SSE was investigated by two different research designs. First design included a post-hoc simulation using responses of real individuals who took paper and pencil format of SSE on past years. Second design includes a live CAT implementation to individuals using a CAT interface and an item bank including item from past SSE.

Data sets used were obtained from Student Selection and Placement Center. In CAT implementation since each individual is given different items, ability estimations obtained should be comparable. For that reason, models of Item Response Theory (IRT) (Embretson & Reise, 2000; De Ayala, 2009) were used since IRT provides ability estimations independent of items. That feature makes comparison of ability levels obtained using different items possible. Another striking feature of IRT is that it provides unique standard error (SE) of ability estimation for each individual. Reliability of scores obtained from CAT implementation is expressed as standard errors of ability estimations. Among the dichotomous IRT models, 3-parameter logistic (3PL) model was employed in the present study. 3PL relates three parameters (item discrimination, item difficulty, and pseudo-guessing factor) to ability level. In paper and pencil format of SSE, a correction formula is applied by deleting 1 true response for

each 4 wrong response. By that way, guessing factor is tried to be eliminated or, at least, lessened. However, IRT analysis revealed that there is still a guessing factor, so 3PL was used. But 2PL, which does not include guessing factor as a parameter also fitted, and can also be utilized.

Items were calibrated and ability estimation of test-takers of SSE science subtest was calculated using BILOG-MG (Zimowski et al., 1996). Details of that phase can be seen on Kalender's study (2011). Computer programs used for that study were developed by researcher and can be demanded via email at no cost.

In both designs, correlation between ability estimations obtained from both simulation and live CAT and paper and pencil format of SSE were compared. Also number of items given by CAT format is another point that was investigated to find out reduction rate in the number of items given.

Post-Hoc Simulation

Post-hoc simulation is a method that depends on using responses of real individuals to past exams. By using computer software, a CAT implementation is simulated for each examinee as if examinee gives the responses that provided in paper and pencil test in a CAT session. By using a post-hoc simulation design, it is possible to find out preliminary information about reduction of items (Weiss, 2011).

Responses of students to SSE 2005 science subtest that includes 45 items were used. To cover different cognitive ability levels, students from three different school types (state, Anatolian, and private high schools) were included to the present study. Mean of science subtest for state, Anatolian, and private high schools are 9.61, 31.72, 24.02 out of 45 items, respectively, which reveals differences of ability levels of students. By using 5,000 randomly selected students for each sample (three school types), ability estimations obtained from post-hoc simulation and those obtained from SSE 2005 science subtest (45 items) were compared.

For ability estimation method for post-hoc simulation, Expected A Priori (EAP), a Bayesian ability estimation method was used. The reason for using EAP is that it produces ability estimation for all response conditions (all correct, all wrong, or combination of true and wrong). Another ability estimation method, Maximum Likelihood that is better to some extent, cannot estimate ability with perfect response patterns (all correct or all wrong) and it requires at least one correct and one wrong response to start ability estimation, whereas EAP does not have such a limitation. Test termination was set on a range of SE between 0.50 and 0.10 with a decrement of 0.10, which correspond to classical reliabilities of 0.75; 0.84; 0.91; 0.96 and 0.99, respectively.

Live CAT

Live CAT included an implementation of CAT format of SSE science subtest to real individuals. 33 university students from English preparatory school of Middle East Technical University were participated to CAT implementation and were given items from an item bank including 242 science items of past SSE. Live CAT session was set to have a SE level of 0.30. To make sure ability estimation if made for all response conditions, Bayesian EAP ability estimation method was used.

Results

Fig. 1 shows the correlations between ability estimates from CAT and paper and pencil format of SSE with respect to school types. As can be seen, the most striking result from post-hoc simulation is that for all school types correlations between ability estimations are higher than 0.80. CAT simulation produced highly correlated scores to paper and pencil format of SSE 2005.

Fig. 1 shows that as SE decreases (reliability increases), correlations also increase. For a SE values of 0.10 all correlations are above 0.95. On the other hand, it is an expected situation that to obtain higher SE values requires more items. Using larger number of items to obtain extremely higher SEs violates one of the principle CAT advantages (fewer items with higher reliable scores). Therefore it is important to note number of items given examinees by CAT required for different SE values.

Table 1 presents medians of numbers of items given to examinees in simulation phase with respect to each SE value. Also amount of reduction in number of items can be seen in parentheses.

For state schools, CAT administration used only 14 and 25 items for a test with SE values of 0.3 and 0.2, respectively. These numbers correspond to a reduction of 68.89% and 44.44% in the number of items given to students. CAT uses much less items compared to paper and pencil format.

Number and percentages of items required for higher degrees of SE is much more than those for lower SEs. For example, to reach a SE of 0.10 numbers of items in item bank are the same with full test length of traditional SSE. All simulation results indicated that to achieve 0.10 degree of SE there is no reduction in the number of items. On the other hand, a SE of 0.30 (equal to CTT reliability of 0.91) can be

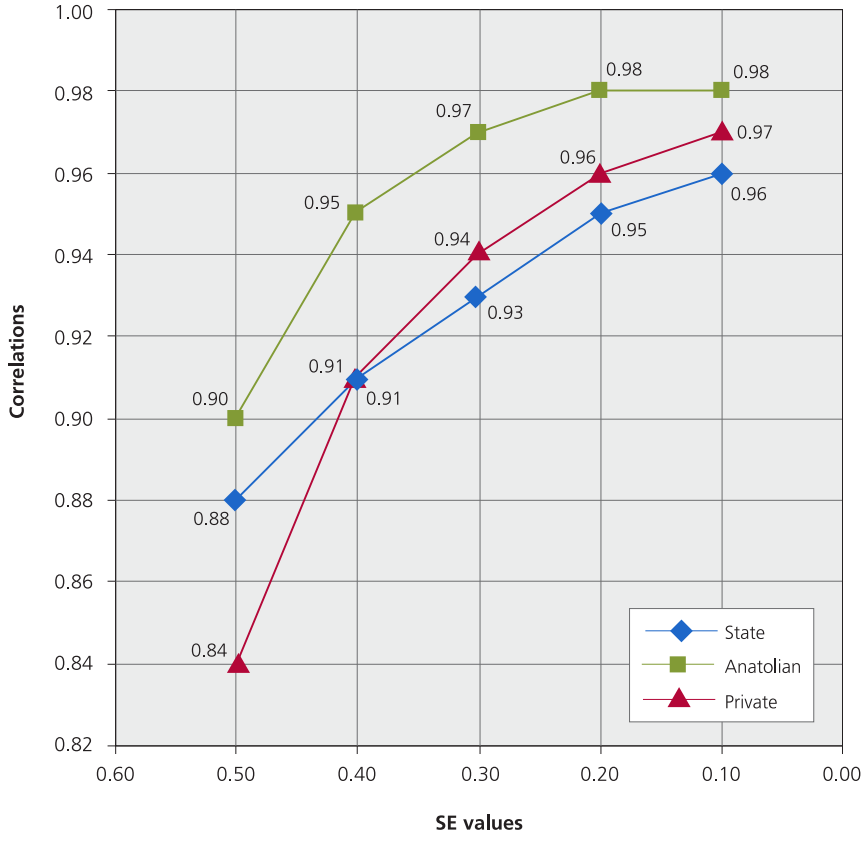

Fig. 1. Correlations between ability estimates

achieved using reasonable number of items. For example, CAT needs 14 items (in median) to achieve 0.30 SE level compared to 45 items given in SSE for state school sample. High correlations obtained for each school types included revealed that school type was not as a factor differentiating ability estimations.

As to live CAT administration, scores of 33 examinees obtained from live CAT and paper and pencil format of SSE has a correlation of 0.74. Correlation between ability estimates can be interpreted as a supporting evidence for applicability CAT implementation of SSE. However a large sample for live CAT research is needed to make generalization. There may be some factors that may lower the correlation found. First, students participated to live CAT administration are university students and they might have seen the questions when they prepared for SSE. Second, students are from a university with high cognitive levels. On the other hand, items are generally suitable for moderate ability groups. Therefore from an item bank with 242 items, computer could not have found proper items for those students.

Table 1. Medians (*percentages*) of number of items used in simulations

Test Length	School Type	Threshold of SE				
		<0.50	<0.40	<0.30	<0.20	<0.10
45	State	6 (86.67)	8 (82.22)	14 (68.89)	25 (44.44)	45 (00.00)
	Anatolian	5 (88.89)	9 (80.00)	15 (66.67)	30 (33.33)	45 (00.00)
	Private	6 (86.67)	12 (73.33)	23 (48.89)	39 (13.33)	45 (00.00)

Median of number of the items given to examinees in live CAT administration phase found to be 9.0 indicating a reduction rate of 80% compared to paper and pencil form of science subtests.

CAT provided a significant reduction in the number of items given to examinees. Also scores of examinees estimated by CAT are highly correlated to scores from SSE. Standard errors, indicators of individual test reliability, are much lower than those obtained from SSE science subtest.

What CAT Provides for Student Selection Examination

Findings of Kalender's (2011) study revealed supporting evidence for applicability of computerized adaptive testing implementation for SSE in Turkey. CAT administration can provide important advantages for SSE such as (not exhaustively):

- First of all, since items are selected dynamically using computer algorithms during the test, individuals are not given items outside the ability levels. Therefore number of items required to estimate ability levels reduces up to 80%. This also shortens testing time.
- Cheating can effectively be prevented by CAT. Since each test-taker receives different items, there cannot be answer copying from near test-takers. Also since there are many items in the item bank and items that will be given to any individual are dynamically determined during the test, problem of reveal of booklets can also be eliminated. Getting help outside testing places via cellular phones, etc. cannot be observed since helpers outside cannot have information that items displayed among the thousands of items.
- Statistical procedures for cheating or collision developed by researchers can be used for CAT implementation. (Wise & Kong, 2005; van der Linden, 2008). Since computers can record any information desired there will be lots of information for each testing session. For example, response time for each item can be used for cheating analysis. If a test-taker gives a response in a significantly short time, that can be trigger for cheating. Also test session can be recorded as a movie clip and if a problem or objection occurs for test results, these clips can be used for validation.

Going Adaptive for Student Selection Examination

To conduct a CAT implementation, item bank should be large enough so as to computer finds items appropriate for whole range ability level of test-takers and SSE is not an exception. Large item bank is also important in that as individuals take test, items in the item bank can quickly reveal or individuals encounter the same items in different sessions if item bank is not large enough. Student Selection and Placement Center use similar types of item and have a large item bank for SSE. It is possible to transform item parameters of those items for using in CAT implementations with minor modifications.

By nature of CAT, test-takers receive different items according to their ability levels and that may cause in a rise of concerns in public. Each test-taker will receive items different in difficulty and number. Placement based on such testing trigger objections as to equality of tests. A transformation like that in test format is not easily acceptable by stake-holders of student selection system. Public should be informed about nature of CAT, equality of scores among test-takers, etc.

Related to different items for different individuals, in February 17, 2011, a law for restructuring duties and organization of Student Selection and Placement Center (also that name has been changed) has been enacted and the law made adaptive format possible in respective section of the law, which says "examinations can be conducted in a way that participants take different items that can be changed according to responses they give and participants can take examinations at different times." (Ministry of National Education, 2011).

Another point that should be discussed is content validity. In paper and pencil format of SSE, for example, science subtest includes physics, chemistry and biology items. In CAT administration, computer selects items from many items and if some restrictions are not applied, items may be selected from just one dimension. To prevent such a possibility, computer algorithms can easily be developed to ensure that test-takers receive items from all subdimensions in pre-specified values.

Before using CAT format for SSE, other large-scale tests such as the Entrance Examination for Graduate Studies can be given in adaptive. Group of test-takers for that exam is graduates or seniors and may more easily accept such changes in test format. As acceptance of CAT by public increases, issues related to CAT administration for SSE could be arisen. CAT format can allowed to be taken more than once in a period of time or CAT format can be optional for test-takers. Both types of SSE can also be given so test-takers can decide which scores they submit for placement. By this way, familiarity for CAT can be obtained among public and also anxiety of students can be lowered.

Transformation of SSE to CAT format is not only a issue of measurement discipline. Development of computer programs, improving network infrastructure, securing online content, development of user-friendly interfaces are some of the topics that are be handled.

As the findings discussed above revealed, CAT administration of SSE yielded highly correlated ability estimations with traditional SSE using fewer items without loss of reliability beside other advantages discussed above. Although findings are related to SSE, other large-scale tests can also administered by CAT format. As a result, the findings given above showed that CAT administration can effectively be used for student selection to higher education programs in Turkey, as well as other large scale test administrations.

References

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

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

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341-349.

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

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

Kalender, I. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. Unpublished doctoral dissertation, Middle East Technical University, Ankara, Turkey.

Mead, A. D., and 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 (2011). *Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığının Teşkilat ve Görevleri Hakkında Kanun*. Accessed through <http://mevzuat.meb.gov.tr/html/27863_6114.html> on January 12th, 2012.

Rudner, L. M. (1998). *An on-line. Interactive computer adaptive testing mini tutorial*. Accessed through <http://edres.org/scripts/cat/catdemo.htm> on December 25th, 2010.

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

Student Selection and Placement Center (2011). *Statistics about scores*. Accessed through <http://www.osym.gov.tr> on June 13rd, 2011.

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

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

Weiss, D. J. (2011). *CAT Central: A global resource for computerized adaptive testing research and applications*. Accessed through <http://www.psych.umn.edu/psylabs/CATCentral> on May 5th, 2011.

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

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