Eurasian Journal of Educational Research, Issue 49, Fall 2012, 61-80

Application of Computerized Adaptive Testing to Entrance Examination for Graduate Studies in Turkey

Okan BULUT*

Adnan KAN**

Suggested Citation:

Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research,* 49, 61-80.

Abstract

Problem Statement: Computerized adaptive testing (CAT) is a sophisticated and efficient way of delivering examinations. In CAT, items for each examinee are selected from an item bank based on the examinee's responses to the items. In this way, the difficulty level of the test is adjusted based on the examinee's ability level. Instead of administering very long tests, CAT can estimate examinees' ability levels with a small number of items. A number of operational testing programs have implemented CAT during the last decade. However, CAT hasn't been applied to any operational test in Turkey, where there are several standardized assessments taken by millions of people every year. Therefore, this study investigates the applicability of CAT to a high-stakes test in Turkey.

Purpose of Study: The purpose of this study is to examine the applicability of CAT procedure to the Entrance Examination for Graduate Studies (EEGS), which is used in selecting students for graduate programs in Turkish universities.

Methods: In this study, post-hoc simulations were conducted using real responses from examinees. First, all items in EEGS were calibrated using the three-parameter item response theory (IRT) model. Then, ability estimates were obtained for all examinees. Using the item parameters and responses to EEGS, post-hoc simulations were run to estimate abilities in CAT. Expected A Posteriori (EAP) method was used for ability estimation. Test termination rule was standard error of measurement for estimated abilities.

* Research Assist.University of Minnesota, Department of Educational Psychology , bulut003@umn.edu

** Assoc Prof Dr. Gazi University, Department of Educational Sciences, adnankan@gazi.edu.tr

Findings and Results: The results indicated that CAT provided accurateability estimates with fewer items compared to the paper-pencil format of EEGS. Correlations between ability estimates from CAT and the real administration of EEGS were found to be 0.93 or higher under all conditions. Average number of items given in CAT ranged from 9 to 22. The number of items given to the examinees could be reduced by up to 70%. Even with a high SEM termination criterion, CAT provided very reliable ability estimates. EAP was the best method among several ability estimates methods (e.g., MAP, MLE, etc.).

Conclusions and Recommendations: CAT can be useful in administering EEGS. With a large item bank, EEGS can be administered to examinees in a reliable and efficient way. The use of CAT can help to minimize the cost of the test since test booklets, examinee response sheets, etc. won't be needed anymore. It can also help to prevent cheating during the test.

Keywords: Computerized adaptive testing, item response theory, standardized assessment, reliability.

Standardized tests in Turkey are implemented in such a way that a multiple-choice test in a paper-pencil format with the same items for everyone is given to all examinees on a certain date. Most of the large-scale assessments in Turkey are administered by the Student Selection and Placement Center and Ministry of National Education. The Student Placement Examination, the Foreign Language Examination for Civil Servants, the Entrance Examination for Graduate Studies, and the Level Determination Exam are some of the high-stakes tests that are taken by many examinees in Turkey every year. For example, over one million examinees take the Student Selection Examination (SSE), which is used for placing students into undergraduate programs in Turkish universities. The Foreign Language Examination for Civil Servants, which is used for measuring English reading comprehension skills of public servants, is also taken by thousands of people. The Entrance Examination for Graduate Studies (EEGS), which is similar to the GRE in the US in terms of its purpose, is taken by fourth-year undergraduates and college graduates. EEGS scores are submitted with graduate school applications in Turkey (Student Selection and Placement Center, 2012).

Among these tests, EEGS is an important one because scores obtained from EEGS are used for admitting students to graduate programs and also for selecting graduate assistants in Turkish universities. One big criticism of EEGS might be the lack of stability in difficulty level and scores that can't be compared from year to year. Since scores for the EEGS subtests are obtained using traditional Classical Test Theory (CTT) methods, test scores depend heavily on items used in the test and persons taking the test. For instance, a person may attend two administrations of EEGS within the same year and obtain very different scores, although the ability level of the person hasn't changed much between the two administrations. This is due to the

fact that test scores and item difficulties are weighted based on the performance of other test-takers in a particular test administration. Therefore, test scores from EEGS can be substantially biased for some examinees.

Another issue with EEGS is the lack of stability in the precision of test scores. Since only a specific set of items is administered to each examinee, it is hard to compute test scores for everyone at a similar level of precision. Also, the use of all items for all examinees may not be necessary because some items may provide very small amounts of information or no information for some examinees with a particular ability level. For example, some items can be very hard or very easy for some examinees. This situation may cause several disadvantages. First, items that are not suitable for an examinee's ability level provide only a little information about the ability level. Second, administering very difficult or very easy items to examinees can make them bored or frustrated. Thus, using such items would be a waste of time. Also, examinees may attempt to guess the answers to items that are very difficult for them, which may, in turn, increase the error inability estimation. If it is possible to give each examinee a test with an ideal matching to his/her ability level, the problems mentioned above could be solved effectively (Mead &Drasgow, 1993).

As described earlier, matching test items with examinees' ability levels is an important issue in all testing programs. To administer items that would match examinees' ability levels, Weiss (1983) suggested using responses for previously given items in the test to select the next appropriate items for an examinee. Computerized adaptive testing (CAT) is a procedure that put this idea into practice. CAT is a special approach to the assessment of latent abilities in which the selection of the test items presented to the examinee is based on the responses given by the examinee to previously administered items (Frey & Seitz, 2011).The basic idea behind CAT is to give examinees only items tailored or adapted to their ability levels in order to maximize the information drawn from each response. In a typical CAT administration, an iterative process with the following steps is used:

1. All the items that have not yet been administered are evaluated to determine which will be the best one to administer next given the currently estimated ability level.
2. The best next item is administered and the examinee responds.
3. A new ability estimate is computed based on the responses to all of the administered items.
4. Steps 1 through 3 are repeated until a stopping criterion is met (Rudner, 2012).

Among the advantages of CAT over conventional testing, Betz and Weiss (1974) stated that CAT-based tests are shorter than conventional form sand provide precise ability estimates of examinees. Embretson (1996) also mentioned that CAT requires fewer items, producing more valid measurement experiences than paper and pencil tests. Another advantage of CAT is its capacity to substantially increase measurement efficiency, which is the ratio of measurement precision totest length

(Frey& Seitz, 2009; Segall, 2005). Compared to conventional testing programs that mostly administer a fixed number of items in a fixed order, CAT can reduce the number of items by approximately half without a loss of information and precision (e.g. Segall, 2005).Although most CATs use item pools that have been calibrated with a unidimensional item response theory (IRT) model (e.g., van der Linden & Hambleton, 1997), there are multidimensional and bi-factor CAT algorithms for tests with a multidimensional structure as well (e.g., Segall, 1996, 2001; Wang & Chen, 2004).

Currently, there are many operational programs that carry out CAT in different fields. Some examples are Graduate Record Examination (GRE) for graduate school admissions and Graduate Management Admission Test (GMAT) for business school admissions in the US, Japanese Computerized Adaptive Test (J-CAT)for diagnosing the proficiency level of Japanese as a second language, Paramedic exams by National Registry of Emergency Medical Technicians for certifying the competency of entry-level emergency medical technicians. Also, a number of testing programs and tests are working toward the implementation of CAT; they include the United States Medical Licensing Examination (USMLE)of the National Board of Medical Examiners (IACAT, 2012).

Comparing the popularity of CAT and the comprehensive literature about its applications in the US and other countries, CAT is still a fairly new area in Turkey. There are only a few studies that examined applicability of CAT to different standardized assessments in Turkey. In an early study, Köklü (1990) compared adaptive and paper-pencil test formats in terms of validity and reliability. Results indicated that there was no statistically significant difference between reliability estimations of the adaptive and conventional formats. However, when the researcher investigated the relationship between test scores from adaptive and paper-pencil formats and students' grades in a science class to study the validity of testing formats, he found correlation coefficients of 0.88 and 0.81 for adaptive and conventional testing formats, respectively. Although differences were not very large, CAT administration provided better results.

Kaptan (1993) conducted a similar study by comparing ability estimates obtained from a conventional paper test and a computerized adaptive test. In the study, examinees took a 50-item math test in paper-pencil format and a 14-item CAT test. Results indicated that CAT provided a 70% reduction rate in the number of items administered. Also, there was no significant difference found between the ability estimates from CAT and the conventional test. Yaşar (1999) investigated KR-20 reliability coefficients of CAT. Correlations obtained from CAT and the paper-pencil format of the same test were compared. In the study, the CAT item bank included 61 items. Correlation between the two formats was found significant with a coefficient of 0.36, indicating a low relationship. The researcher indicated some potential reasons for that, such as limited number of items in the bank and a test stopping rule with fixed number of items. In a similar study, Iseri (2002) constructed an item bank using the items in the Secondary School Student Selection and Placement Examination. Iseri (2002) stated that CAT estimated students' achievement levels

using fewer items. In test sessions in which students were allowed to go back to the items responded to earlier, estimations for students with higher ability level was better than those with lower levels. The Bayesian estimation method provided better ability estimates. Also, both of the stopping rules (fixed number of items and fixed standard error) yielded reliable results.

Kalender (2011, 2012) applied computerized adaptive testing tothe science subtest of Student Selection Examination in Turkey. A post-hoc simulation study and a live CAT study were conducted. Expected A Priori (EAP) was used for estimating abilities, with standard errors ranging from .10 to .50 as test termination criteria. Results showed that CAT provided a reduction by up to 80% in the number of items given to students compared to the paper and pencil form of the test. Correlations between ability estimates obtained from CAT and the full-length test were higher than 0.80.For the live CAT administration, this correlation was about.74, which might be due to the small sample size (33 persons) used in the study. After recent cheating issues in standardized assessments in Turkey, Kalender (2012) argues that the use of CAT can help to prevent cheating since each person receives different items during the test.

More research is needed to examine the applicability of CAT to different testing programs in Turkey. CAT can be a solution to the current issues with the high-stakes tests in Turkey. The Entrance Examination for Graduate Studies is an exam that CAT can be applied to more easily. As Kalender (2012) mentioned, transition from the conventional testing to CAT can be relatively easier for EEGS because persons eligible to take EEGS are mostly college graduates who are used to different test formats. Therefore, they can more easily adapt themselves to such a change in test format more easily. This study applies CAT to the Entrance Examination for Graduate Studies (EEGS) in Turkey and shows the benefits of this method over the paper-pencil testing. The purpose of the study is to compare ability estimates from CAT and paper-pencil administrations results through a post-hoc simulation study by using different ability estimations and test termination criteria.

Method

Research Design

The purpose of this study is to examine applicability and efficiency of CAT for the subtests of EEGS. Through post-hoc simulations, performance of CAT will be compared to the conventional (i.e. paper-pencil format) testing. There are two research questions for this study:

1) How does the CAT perform for estimating ability levels of examinees in EEGS compared to the conventional paper-pencil format?
2) Do different test termination conditions (i.e., SEM) affect ability estimation and test length during the CAT administration?

A post-hoc simulation method was used to examine applicability of CAT for EEGS. The post-hoc approach to simulation is used when CAT is to be used to reduce the length of a test that has been administered conventionally (Weiss, 2012).In this approach, the item bank for CAT consists of all items administered to test-takers in the test. This type of simulation study can help to determine how much reduction in test length can be achieved by re-administering the items in an adaptive way without changing the psychometric properties of the test scores.

Sample

The data for this study come from the 2008administrations of the Entrance Examination for Graduate Studies (EEGS).Results of EEGS are used for admitting students to graduate programs and selecting graduate assistants in Turkish universities. Fourth-year undergraduate students and college graduates are eligible to take the test. The test is administered twice a year in a conventional form (i.e. paper-pencil test).EEGS consists of three subtests: quantitative 1, quantitative 2, and verbal. Each of the quantitative 1 and quantitative 2 sections has 40 items that measure mathematical and logical reasoning abilities. The quantitative 2 section has more advanced and difficult items than does quantitative 1. The verbal section has 80 items that measure verbal reasoning ability. All items in EEGS have five response options and they are scored dichotomously.

To conduct a post-hoc CAT analysis, a random sample of 10,000 examinees (5,000 male, 5,000 female) was selected from the full dataset. The sample includes examinees from 123 universities in Turkey and outside Turkey. Examinees' ages ranged from 18 to 61.Table 1 shows the summary statistics for the total scores from EEGS.

Table 1

Summary Statistics for the Total Scores in the Three Subtests of EEGS

Test	# of items	Alpha	Mean	SD	Min	Max
Quantitative 1	40	0.96	23.28	11.92	0	40
Quantitative 2	40	.97	18.36	13.31	0	40
Verbal	80	.96	59.72	16.66	0	80

Data Analysis

In this study, the post-hoc simulation procedure described by Weiss (2012) was used:

1. Item parameters based on an item response theory (IRT) model are estimated using the available item response data.
2. Then, using these item parameters, abilities (theta) are estimated for each examinee.
3. A test termination criterion (e.g., a standard error of .3 or fixed number of items) is determined.
4. The CAT is implemented by selecting items adaptively for each examinee and the CAT is terminated based upon the pre-specified termination rule.

5. Final theta values are estimated for each examinee using maximum likelihood (MLE) or Bayesian methods.
6. The CAT theta estimates are compared with the conventional test theta estimates based on the number of items administered in the CAT.

By following these steps, first, item parameters for quantitative 1, quantitative 2, and verbal subtests of EEGS were estimated using the three-parameter logistic IRT model (3PL) in Xcalibre 4.1 (Guyer & Thompson, 2011).IRT model assumptions (i.e., unidimensionality and local item independence) have been checked for the subtests of EEGS. All three subtests were found appropriate for IRT modeling. The 3PL model has the best model-data fit for EEGS among other unidimensional IRT models (Bulut, 2010). The 3PL unidimensional IRT model can be shown as follows:

$$P\{X_{ij}=1|\theta_i,a_j,b_j,c_j\}=c_j+(1-c_j)\frac{\exp\left[a_j(\theta_i-b_j)\right]}{1+\exp\left[a_i(\theta_i-b_j)\right]} \quad (1)$$

where θ_i is the unidimensional ability estimate for person *i*, b_j is item difficulty for item *j*, a_j is item discrimination for item *j*, and c_j is guessing parameter for item *j*. Summary statistics for the calibrated items and summary statistics for the ability estimates for the three subtests of EEGS are presented in Table 1 and Table 2, respectively. Also, test information functions (TIF), which show the information and precision of items in the test, and standard error of measurement based on the 3PL model for each subtest of EEGS are shown in Figure 1.

Table 1

Summary Statistics for all Calibrated Items in the Three Subtests of EEGS

Test	Parameter	Mean	SD	Min	Max
Quantitative 1	a	2.450	0.607	1.189	3.735
	b	-0.089	0.484	-0.905	1.107
	c	0.089	0.053	0.036	0.268
Quantitative 2	a	3.066	0.790	1.490	4.183
	b	0.289	0.459	-0.621	1.541
	c	0.046	0.025	0.021	0.157
Verbal	a	1.993	1.179	0.438	4.128
	b	-1.281	1.217	-3.848	0.654
	c	0.039	0.026	0.021	0.119

Table 2

Summary statistics for the ability estimates from the three subtests of EEGS

Test	Max Info	Min CSEM	Mean	SD	Min	Max
Quantitative 1	42.574	0.153	0.004	0.990	-2.120	1.981
Quantitative 2	72.343	0.118	0.001	1.010	-1.699	2.169
Verbal	57.807	0.132	0.006	1.007	-3.853	2.001

Note: CSEM = Conditional standard error of measurement.

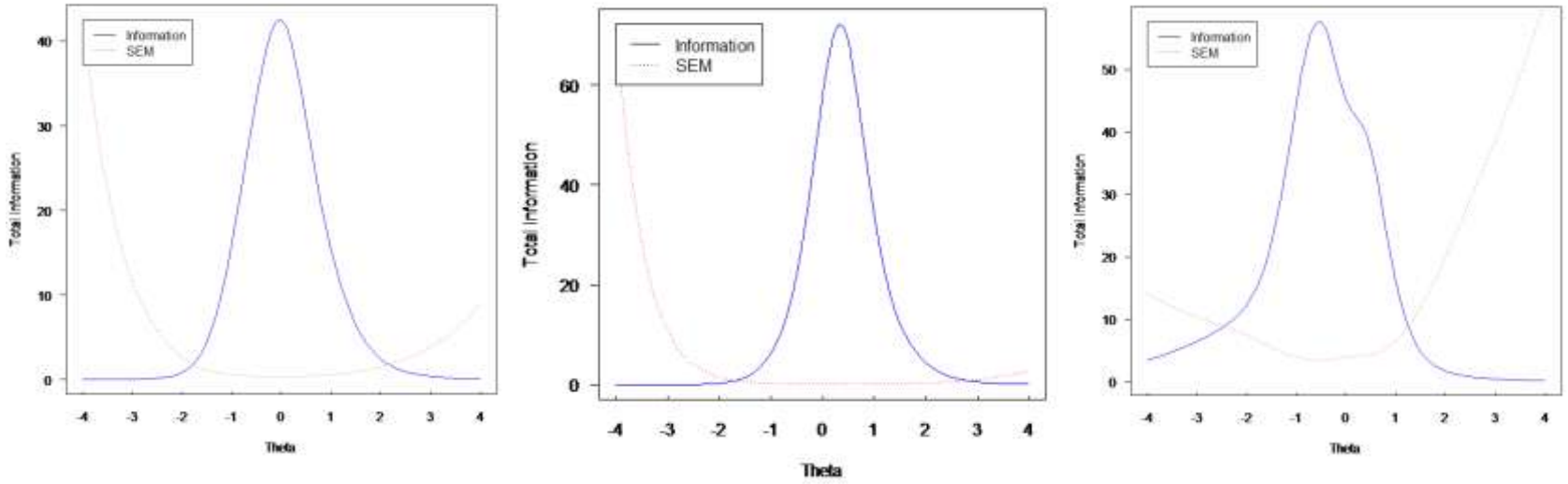

*Figure 1.*Test information function and standard error of measurement for quantitative 1 (left), quantitative 2 (middle), and verbal(right) subtests

After item parameters were obtained, theta (θ) values based on Expected a Posteriori (EAP) method were estimated for all examinees using the same software. EAP estimator was preferred in this study because, unlike the Maximum Likelihood (ML) estimator, EAP does not rely on an iterative procedure and uses a closed form estimator (i.e., a simple integration using numerical quadrature). Another advantage of EAP over ML is that it provides a finite estimate for the perfect and null scores. Thus, EAP can provide a finite estimate after the first item, even if the response was in one of the two extreme categories (Choi, Podrabsky& McKinney, 2010). Although EAP was used for estimating abilities and computing all accuracy measures, ability estimates from Maximum Likelihood (MLE), Maximum a Posteriori (MAP), and Weighted Least Square (WLS) were also obtained to look at the relationship between EAP and other ability estimators.

In the next step, estimated item parameters and person abilities were used to configure a CAT administration. Firestar-D (Choi, 2009; Choi et al., 2010) was used for running post-hoc CAT analyses. Firestar-D generates R codes (R Development Core Team, 2012) for implementing post-hoc CAT analyses based on pre-specified item selection and test termination criteria. In this study, the maximum Fisher information (MFI) method was used as item selection method. MFI method can be shown as follows:

$$i_k = \arg\max_j \{I_j(\hat{\theta}_{u_1,u_2,\ldots,u_{k-1}}) : j \in R_k\} \quad (2)$$

The MFI method iteratively selects the next item that provides maximum information at a particular $\hat{\theta}$. Every selected item provides the greatest increase in test information and the greatest reduction in standard error.CAT can be terminated when each examinee is measured with a pre-specified degree of precision, which allows measurement of θ levels of all examinees equally. In several test settings, CAT is terminated when a predetermined number of items is reached. However, using a fixed number of items as the termination criterion may be inappropriate for CAT because it does not provide all examinees with equal precision in measuring θ (Weiss, 2004). In this study, a fixed standard error of measurement (SEM) was used as the termination criterion for the CATs in the post-hoc simulations. The test is terminated when SEM for the estimated theta estimate drops below the pre-specified SEM value. A number of SEM termination criteria (.25, .30, and .40) were used for each subtest of EEGS. Figure 2 shows a visual example of this iterative process.

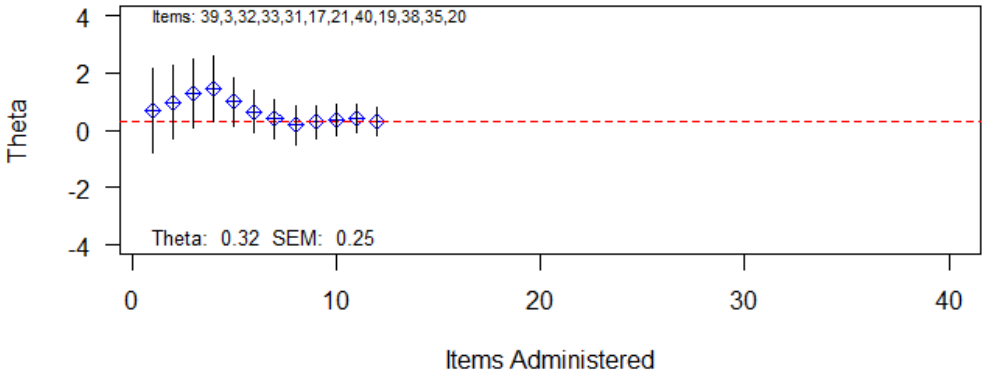

Figure 2. An example of the adaptive ability estimation process of an examinee in CAT

After post-hoc CAT simulations were completed for the three subtests of EEGS, the following evaluation criteria from Weiss and Gibbons (2007) were computed to compare the performance of CAT to the conventional testing of EEGS:

1. The average number of items required by CAT to recover full-scale θ estimates with a pre-specified standard error of measurement.
2. Pearson correlations between CAT θ estimates ($\ddot{\theta}_C$) and full-scale θ estimates ($\hat{\theta}_F$).
3. Average signed difference (i.e. bias)between CAT and full-scale θ estimates:

$$Averaged\ signed\ difference = \frac{\sum_{i=1}^{N}(\ddot{\theta}_C - \ddot{\theta}_F)}{N}$$

4. Average absolute difference(i.e. accuracy) between CAT and full-scale θ estimates:

$$Average\ absolute\ difference = \frac{\sum_{i=1}^{N}|\ddot{\theta}_C - \ddot{\theta}_F|}{N}$$

5. Root mean squared difference (RMSD) between CAT and full-scale θ estimates:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_C - \hat{\theta}_F)^2}{N}}$$

Results

Post-hoc CAT simulations were implemented using the item parameters, theta estimates, and item responses from the full-length test as described above. Table 3 presents the results of post-hoc CAT simulations for each subtest of EEGS. The results showed that CAT was able to recover abilities accurately under all SEM conditions for each subtest. The correlation between the ability estimates from the full-length test and the CAT administration was .93 or higher for all subtests. These results indicated that CAT ability estimates are aligned with the abilities from the full-length test. CAT significantly reduced the number of items administered to the examinees. The reduction rate ranged from 44% to 88%. The highest reduction rate was observed in the verbal subtest. The correlation between the CAT ability estimates and the abilities from the whole test changed depending on the SEM termination rule. As SEM increased, the correlation between CAT abilities and full-test abilities decreased. On the contrary, reduction in the number of items administered increased as SEM for test termination increased. Figure 3 shows the relationship between the number of items administered and ability levels when SEM was 0.25.

Table 3

Correlation Between theta Values from CAT and the Full Test, Bias, Accuracy, Mean, and Range of Number of Items Administered

					Number of items		
Subtest	SEM	$r(\hat{\theta}_C, \hat{\theta}_F)$	Bias	Accuracy	Mean	Range	Reduction
Quantitative 1	.25	.98	-.004	.089	22.39	8-40	44%
	.30	.97	-.009	.129	17.50	7-40	56%
	.40	.95	-.012	.217	11.15	4-40	72%
Quantitative 2	.25	.98	.010	.105	19.88	6-40	50%
	.30	.97	.016	.134	16.60	5-40	59%
	.40	.94	.036	.204	11.95	4-40	70%
Verbal	.25	.96	.016	.152	22.11	8-80	72%
	.30	.95	.031	.187	15.20	6-80	81%
	.40	.93	.036	.249	9.05	5-80	88%

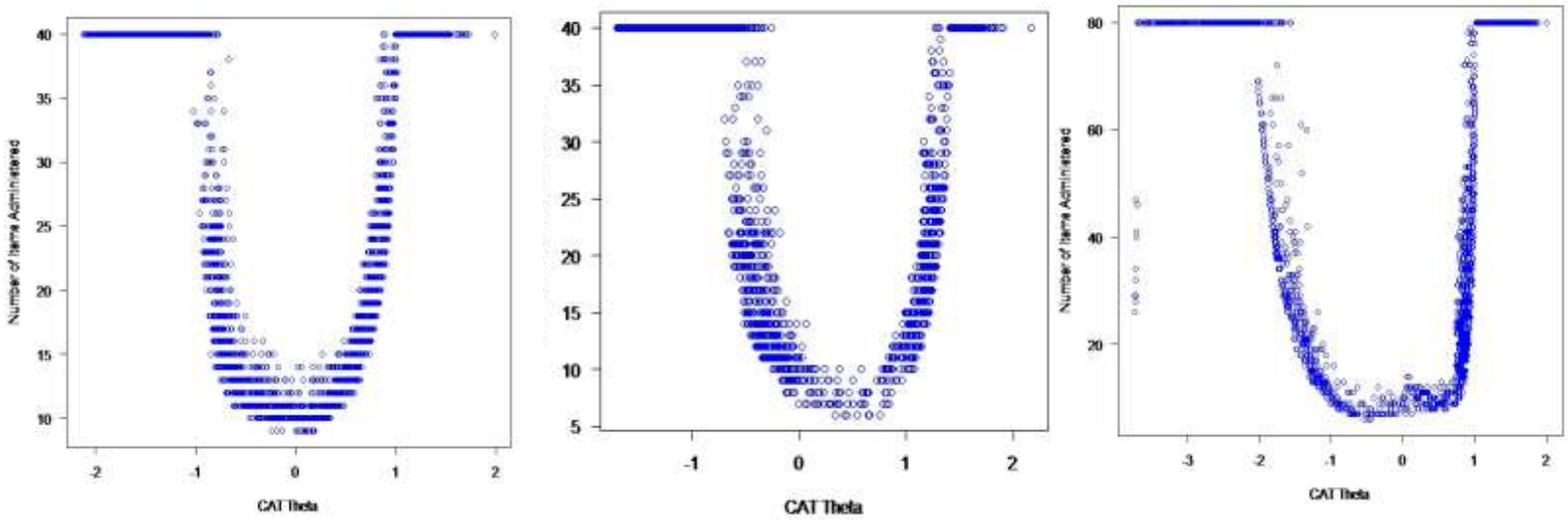

Figure 3. Number of items administered at different theta levels for quantitative 1 (left), quantitative 2 (middle), and verbal (right) subtests when SEM = .25

Bias for all subtests was negligible. There was a negative bias in the ability estimates for the quantitative 1, whereas there was a positive bias in the ability estimates for quantitative 2 and verbal subtests. Verbal subtest had the highest bias, although this subtest had more items than the others. Also, the verbal subtest had the lowest accuracy among the three subtests. The reason for this result was that the verbal subtest failed to estimate extreme abilities (i.e., very low or high) accurately despite having more items. Since the number of items for each subtest was very limited, the items were not able to cover all ranges of abilities. Therefore, each subtest was able to measure only a certain level of abilities accurately. For the examinees with very high or low ability levels, SEM test termination criterion wasn't met, even when all items were administered. Similar to bias, RMSD also increased as the SEM value for test termination increased (see Figure 4). The verbal subtest had the largest RMSD among the three subtests under each of the SEM-based test termination criteria. Based upon these results, the CAT carried out the most accurate ability estimation for the quantitative 1 subtest and the least accurate ability estimation for the verbal subtest.

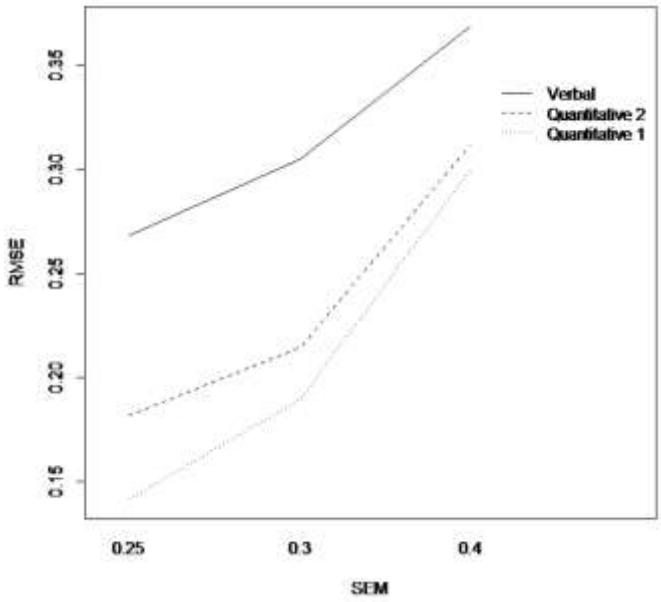

Figure 4. Change of RMSD based on the amount of SEM for the three subtests of EEGS

As described earlier, EAP method was used for estimating abilities in the post-hoc CAT simulations. In addition to EAP, Maximum Likelihood (MLE), Maximum A Posteriori (MAP), and Weighted Least Squares (WLS) methods were used to estimate the final ability estimates from the CAT administrations. Table 4 shows the correlation between CAT-based EAP abilities and other abilities obtained from MLE,

MAP, and WLS methods. As seen in Table 4, EAP and MAP estimates were always highly correlated. MLE and WLS estimates were also highly correlated with EAP estimates. However, especially for very high or very low abilities, MLE and WLS methods were not able to recover the abilities as accurately as the EAP estimator. Since the regular MLE fails to estimate persons with completely wrong or completely correct responses, which is commonly observed in EEGS, the EAP estimator can be more appropriate for estimating persons' abilities.

Table 4

Correlations Between Ability Estimates from EAP and Other Estimators in CAT

Test	SEM	$r_{(EAP, MAP)}$	$r_{(EAP, MLE)}$	$r_{(EAP, WLS)}$
	.25	.99	.95	.96
Quantitative 1	.30	.99	.95	.96
	.40	.99	.95	.96
	.25	.99	.93	.95
Quantitative 2	.30	.99	.93	.95
	.40	.99	.94	.95
	.25	.99	.98	.98
Verbal	.30	.99	.98	.98
	.40	.99	.98	.97

Conclusion and Discussion

This study examined the applicability of computerized adaptive testing (CAT) to the Entrance Examination for Graduate Studies (EEGS) in Turkey. Using real examinee responses from the 2008 administration of EEGS, a series of post-hoc CAT analyses were carried out. EAP was used for estimating abilities during the CAT. A fixed standard error of measurement (SEM) was used for terminating the CAT. Post-hoc simulations provided results supporting the applicability of CAT administration in EEGS. CAT was able to recover persons' abilities precisely with many fewer items than the full-length form of EEGS. Although the examinees with very high or low ability levels still had to take all items in the test, the rest of the examinees were measured with a smaller number of items and high precision. EAP estimator seemed to be a better estimation method for EEGS compared to other methods (e.g. MLE and WLS). Since this was a real CAT implementation, the items in the test were informative only within a specific range of abilities. Therefore, CAT provided more precise measurement for examinees within that range than examinees with extreme abilities.

Developing an item bank would be the most important part of CAT implementation for EEGS. To provide equiprecise measurement, which means measuring everyone with the same level of precision, item bank should have a sufficient number of test items properly distributed across the theta scale and the CAT should be allowed to continue long enough for each examinee (i.e., no fixed number of items as a termination rule). As Kalender (2011) also stated, the item bank should be large enough so that the CAT algorithm can pick the most appropriate items for test-takers with different levels of ability. Therefore, the item bank for CAT should have a number of high-quality items to increase the efficiency of CAT.

With a high-quality item bank, CAT can significantly reduce the time spent on responding items. Since EEGS is a long test, examinees may get bored during the exam and start making random guessing or skipping items. Instead of administering the whole test in a conventional form, CAT can provide the most appropriate items from the item bank for each examinee and reduce the testing time. In this way, the problems of random guessing and skipping numerous items can be minimized. CAT can also reduce the cost of the exam. Every year hundreds of thousands of test booklets and answer sheets are printed for EEGS. In addition to printing costs, transportation and securing of these testing materials cause additional costs. The use of CAT can allow administering EEGS several times within a year without printing hundreds of thousands of test booklets. CAT would also be an important convenience for persons who plan to take the test since they would not feel under pressure for taking the test on a certain date and time.

Implementation of CAT is also useful for detecting persons who attempt to cheat on the test. First, since all responses are saved in a computer, there is no way to steal test booklets before or during the test. Also, there are several statistical procedures developed for detecting cheating or unexpected response behaviors on the test (e.g. Wise &Kong, 2005; van der Linden, 2008). Response times or response patterns can be used for investigation of cheating. Very short response times or unexpected response patterns might be an indicator of cheating. In most operational CAT programs such as GRE by Educational Testing Service (ETS), a camera records the entire session in the testing room. In case of suspicious responding behaviors, these recordings can be examined to find the problem.

This study had some limitations. First, since this was a post-hoc study, there were only a limited number of items in the item bank. Therefore, the item bank was able to cover only a certain range of abilities. An item bank with more items is needed to better test the performance of CAT for EEGS. A live CAT administration can be carried out with a larger item bank to investigate the performance of CAT administration in a real testing environment. Second, there were no constraints on or balancing of the content in this study. The CAT software picked the most informative item for each person regardless of its content. In a real CAT administration, one may want to pre-specify the number of items to be administered from each content area (e.g., algebra, geometry, etc. in the quantitative sections).

References

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

Bulut, O. (2010). *The fit of one-, two- and three-parameter item response theory models to the Entrance Examination for Graduate Studies in Turkey*. Unpublished master's thesis, University of Minnesota, Minneapolis, MN, USA.

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous IRT models. *Applied Psychological Measurement*, *33*, 644-645.

Choi, S. W., Podrabsky, T., & McKinney, N. (2010). Firestar-D: Computerized adaptive testing Simulation program for dichotomous IRT models (Version 1.4.0) [Software]. Northwestern University, Feinberg School of Medicine.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*, 89-94.

Guyer, R., & Thompson, N.A., (2011).*User's Manual for Xcalibre 4.1*.St. Paul MN: Assessment Systems Corporation.

IACAT - International Association for Computerized Adaptive Testing (2012). Retrieved on 05/31/2012 from http://iacat.org/.

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

Kalender, I. (2012).Computerized adaptive testing for student selection to higher education.*Journal of Higher Education, 2*(1), 13-19.

Kaptan, F. (1993).*Yetenek kestiriminde adaptive (bireysellestirilmiş) test uygulamasi ile Geleneksel kağıt-kalem testi uygulamasının karşılaştırılması* [A comparison of adaptive and conventional paper-pencil testing applications for ability estimation].Unpublished doctoral dissertation, Hacettepe University, Turkey.

Koklu, N. (1990). *Klasik test teorisine gore geliştirilen tailored test ile grup testi arasında bir karşılaştırma* [A comparison between tailored and group tests based on classical test theory]. Unpublished doctoral dissertation. Hacettepe University, Turkey.

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

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rudner, L. (2012). An online, interactive, computer adaptive testing tutorial. Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm on 05/31/2012.

Segall, D. O. (1996).Multidimensional adaptive testing. *Psychometrika*, *61*, 331-354.

Segall, D. O. (2001). General ability measurement: An application of multidimensional itemresponse theory. *Psychometrika, 66*, 79-97.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York, NY: Academic Press.

Student Selection and Placement Center. (2010). Retrieved on from http://www.osym.gov.tr05/31/2012.

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

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency ofmultidimensional computerized adaptive testing. *Applied Psychological Measurement,28*, 450-480.

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. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37,* 70-84.

Weiss, D. J. (2012).*CAT Central: A global resource for computerized adaptive testing research and applications*. Retrieved fromhttp://www.psych.umn.edu/psylabs/CATCentralon 05/31/2012.

Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing,* URL. http://www.psych.umn.edu/psylabs/catcentral/pdf files/cat07weiss&gibbons.pdf

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.

Yaşar, M. (1999). *Bireysellestirilmiş testler üzerine bir çalışma* [A research study on adaptive testing]. Unpublished doctoral dissertation, Hacettepe University, Turkey.

Bilgisayar Ortamında Bireyselleştirilmiş Testlerin Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'na Uygulanması

Atıf:

Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, 49, 61-80.

(Özet)

Problem Durumu

Son yıllarda dünya genelinde yaygınlaşmaya başlayan bilgisayar ortamında bireyselleştirilmiş (CAT) test uygulamaları halen kullanılmakta olan klasik testlere göre çok daha güvenilir ve hızlı sonuçlar alınmasını sağlamaktadır. Bilgisayar ortamında gerçekleştirilen bu sınavlarda, sınava giren kişiler önceden hazırlanmış bir soru havuzundan kendileri için seçilen sorulara yanıt vermektedirler. CAT sisteminde eğer kişinin her bir soruya verdiği cevap doğru ise bir sonraki soru için soru havuzundan daha zor bir soru, eğer yanlış ise daha kolay bir soru gönderilmektedir. Böylece test kişinin bilgi yada yetenek düzeyine göre ayarlanmış olur. CAT sistemi kullanılan sınavlarda klasik sınavlara göre çok daha az soru ile sınavı alan kişinin puanı güvenilir bir şekilde hesaplanabilmektedir. Çünkü klasik test uygulamalarında olduğu gibi kişi sınavdaki tüm sorulara cevap vermek yerine, kendi bilgi yada yetenek düzeyine uygun olan ve bireyin potansiyelinin en az hata ile kestirilmesini sağlayacak sorularla karşılaşmaktadır.

Türkiye'de her yıl öğrenci seçme ve yerleştirme merkezi ve Milli Eğitim Bakanlığı tarafından birçok sınav düzenlemekte ve bu sınavların sonuçlarına göre üniversite programlarına yerleştirme, devlet memurluğuna atama gibi önemli kararlar verilmektedir. Bu sınavı alan kişilerin bilgi, beceri yada yetenek düzeylerinin en iyi şekilde saptanması büyük önem taşımaktadır. Şuan uygulanmakta olan klasik test yöntemlerine göre CAT sistemi çok daha hızlı ve güvenilir sonuçlar sağlayabilir. Fakat CAT uygulamasına geçilmeden önce eldeki sınavların bu sisteme uygunluğu detaylı bir şekilde araştırılmalıdır.

Araştırmanın Amacı

Bu çalışmanın amacı bilgisayar ortamında bireyselleştirilmiş (CAT) test yönteminin Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'na (ALES) uygunluğunu incelemektir. ALES, yükseköğretim kurumlarında öğretim görevlisi, okutman, araştırma görevlisi, uzman, çevirici ve eğitim öğretim planlamacısı kadrolarına açıktan veya öğretim elemanı dışındaki kadrolardan naklen atamalarda, lisansüstü eğitime girişte, yurt dışına lisansüstü eğitim için gönderilecek adayların seçiminde ilgili kurumların kullanacakları puanları veren bir sınavdır. Bu çalışmada öncelikle CAT sistemi ALES' üzerinde uygulanmıştır. CAT sisteminden elde edilen sonuçlar

ALES sınavının klasik formatta gerçekleştirilmiş halinden elde edilen sonuçlarla kıyaslanmakta ve CAT sisteminin hangi koşullar altında en iyi sonuçlar verdiği tartışılmaktadır.

Araştırmanın Yöntemi

Bu çalışmada ALES'in CAT ve şuan kullanılmakta olan klasik formatlarından elde edilen yetenek kestirimlerini karşılaştırmak amacı ile post-hoc simülasyonlar uygulanmıştır. 2008 yılında uygulanmış olan ALES verileri kullanılarak sınav eğer bilgisayar ortamında CAT sistemi ile gerçekleştirilseydi nasıl sonuçlar elde edilirdi sorusunun yanıtı aranmaktadır. Sınava tüm katılanlar arasından rastgele on bin kişilik bir örneklem seçilmiştir. Bu kişilerin sorulara verdiği cevaplar kullanılarak 3 parametreli madde-cevap kuramı (IRT) modeline göre soruların zorluk ve ayırıcılık indeksleri ve de katılımcıların IRT ölçeğine göre test puanları belirlenmiştir. Sonrasında eldeki sorular bir soru havuzu olarak kullanılarak katılımcıların test puanları bu sefer CAT sistemi ile hesaplanmıştır. Yetenek kestirim yöntemi olarak Expected A Posteriori (EAP) kullanılmıştır. Test sonlandırma kuralı ise standart hata eşik değeri olarak belirlenmiştir. CAT, ALES'in her bir alt testine (sayısal 1, sayısal 2 ve sözel) ayrı ayrı uygulanmıştır. Elde edilen katılımcıların tüm teste verdikleri cevaplardan elde edilen asıl puanları ile karşılaştırılmıştır. Bu karşılaştırmalar için korelasyon ve RMSE gibi indeksler hesaplanmıştır. Post-hoc simulasyonları gerçekleştirmek için Firestar-D programı kullanılmıştır.

Araştırmanın Bulguları

Post-hoc simülasyon bulguları CAT uygulamasının ALES için Expected A Posteriori yetenek kestirim yöntemi ile 0.25, 0.30 ve 0.40 standart hata eşik değeri ile uygulanabileceğini göstermiştir. CAT ve klasik formattan elde edilen yetenek kestirimleri arasındaki korelasyon 0.93 ve üzeri olarak bulunmuştur. CAT ile kullanılan soru sayısı ortalaması ise her bir alt test için 9 ile 22 arasında değişmektedir. Bu sonuçlara göre CAT sistemi ALES' deki soru sayısında yüzde 70'lere varan oranda azalma sağlarken en az tüm sorular uygulandığındaki kadar net yetenek kestirimi sağlamıştır. EAP yetenek kestirim yöntemi ALES için en uygun yöntem olarak görülmüştür. Sayısal 1, sayısal 2 ve sözel alt testleri arasında en fazla hata miktarı sözel testte görülmüştür. Her ne kadar soru sayısı diğer iki alt teste göre daha fazla olsa da soruların sadece belirli bir yetenek aralığını ölçmesinden dolayı çok yüksek ya da düşük yetenekteki katılımcıların puanlarının hesaplanmasında hata oranının yüksek olduğu belirlenmiştir. Sayısal 1 testi normalin biraz daha altında yetenek kestirimleri verirken (negatif yanlılık) sayısal 2 ve sözel alt testleri normalin biraz üstünde yetenek kestirimleri (pozitif yanlılık) sağlamaktadır.

Araştırmanın Sonuçları ve Önerileri

Bu araştırmanın sonuçları bilgisayar ortamında bireyselleştirilmiş test (CAT) sisteminin ALES'e uygulanmasının mümkün olduğunu, uygulandığı takdirde güvenilir sonuçlar sağlayabileceğini göstermektedir. CAT ile yüksek standart hata eşik değeri kullanıldığında bile güvenilir ve net sonuçlar elde edilmektedir. Yeterli

genişlikte bir soru havuzu hazırlanması halinde CAT, sınava giren kişileri sınavın klasik formatındaki kadar çok sayıda soruya tabi tutmadan yetenek kestirimi yapabilmektedir. Bu nedenle CAT'in ALES'e uygulanması aşamasında ilk olarak iyi sorulardan oluşan kaliteli bir soru havuzu oluşturulmalıdır. CAT'in yapacağı bir diğer katkı ise sınavın maliyetini ve değerlendirme süresini düşürecek olmasıdır. CAT ile test kitapçıkları ve cevap formlarının kullanımına gerek kalmamaktadır. Ayrıca her yanıt sonrası yetenek kestirimi yapıldığı için katılıcılar sınav sonrası hemen puanlarını öğrenebilmektedirler. CAT sisteminin kullanılması sınav esnasında kopya çekilmesini de neredeyse imkansız kılacağı için daha güvenilir bir test uygulama süreci sağlamaktadır.

Anahtar Sözcükler: Bilgisayarda bireyselleştirilmiş testler, ALES, madde-tepki kuramı, standart başarı testi.