Eurasian Journal of Educational Research, Issue 58, 2015, 27-44

Ability Level Estimation of Students on Probability Unit via Computerized Adaptive Testing

Hacer ÖZYURT*
Özcan ÖZYURT**

Suggested Citation:
Özyurt, H. & Özyurt, Ö. (2015). Ability Level Estimation of Students on Probability Unit via Computerized Adaptive Testing. *Eurasian Journal of Educational Research*, *58*, 27-44.http://dx.doi.org/10.14689/ejer.2015.58.7

Abstract

Problem Statement: Learning-teaching activities bring along the need to determine whether they achieve their goals. Thus, multiple choice tests addressing the same set of questions to all are frequently used. However, this traditional assessment and evaluation form contrasts with modern education, where individual learning characteristics are featured. Hence, the use of Computerized Adaptive Testing (CAT) systems, which set the difficulty level in accordance with the ability levels of individuals, is spreading. However, these systems are not prevalent in Turkey. Therefore, it is important to develop and assess a CAT system to be integrated into Turkish curriculum.

Purpose of Study: The purpose of this study is to develop a CAT system that can be used in the sub-levels of Turkish curricula (high school) and assess it in terms of reliability.

Methods: In this study, a CAT system aimed at the sub-topics (permutation, combination, binomial expansion, and probability) of the unit of probability covered in 11th grade mathematics was developed. Estimation sensitivity of the developed CAT system, from lower to higher ability levels, was assessed. To this end, an exam was implemented in a high school located in Trabzon in the 2011-2012 fall semester. Eighty-four 11th graders participated in the implementation. Using the data derived from the implementation, reliability coefficient values of each learner were calculated. System records were used for data collection.

Results: Reliability coefficient values for each student in 5 different exams - permutation, combination, binomial expansion, probability, and end-of-unit - were calculated. Findings of the study indicated that the developed

* Dr. Department of Software Engineering, Karadeniz Technical University, Trabzon. e-mail: hacerozyurt@ktu.edu.tr
** Dr. Department of Software Engineering, Karadeniz Technical University, Trabzon. e-mail: oozyurt@ktu.edu.tr

CAT system produced tests with high reliability for all subjects. Average reliability coefficient values for each subject were found as 0.93, 0.93, 0.88, 0.93, and 0.91.

Discussion and Conclusion: By addressing questions from lower to higher levels, which are appropriate to the ability level of each learner, CATs increase sensitivity and reliability of measurement. The developed CAT system addresses questions appropriate for the ability level of the learner, which increases sensitivity in measurement. In addition, it was observed that CAT systems developed for exams have the characteristic of making sensitive measurements ranging from lower to higher levels. This study, in relation to UZWEBMAT-CAT assessment, proved that CAT systems can safely be used within Turkish curriculum.

Keywords: Computerized adaptive testing, individual assessment, individual differences, ability level estimation, adaptive testing.

Learning-teaching activities bring along the need to determine whether they achieve their goals. This need is fulfilled through assessment and evaluation activities (Baki, 2008). Therefore, assessment and evaluation has become an inseparable part of the educational process. Various methods are used for assessment and evaluation in education. One of these methods is multiple-choice tests. These tests are very useful for large-scale exams in particular. Like face-to-face education, assessment is also needed in web-based education, as in conventional educational methods (Semerci & Bektaş, 2005). The advantages brought by internet technology make it possible to simultaneously test many students in a computer environment. However, the traditional assessment and evaluation form contrasts with modern education, where individual learning characteristics are featured. This situation has accelerated the transformation of traditional tests into adaptive tests, which are more suitable for modern education.

Computerized adaptive testing

Computerized adaptive testing refers to implementing different questions to every participant based on their ability levels (Liu & Chen, 2012). In computerized adaptive testing, difficulty of the test dynamically varies by the ability level of the participant. If the participant answers the item correctly, the next item selection prefers a more difficult item. If the participant answers the item incorrectly, the next item selection prefers an easier item (Weiss, 1985). While selecting the items, the ones providing the most information about the ability level of the participant are selected (Bejar, Weiss & Gialluca, 1977; Georgiadou, Triantafillou & Economides, 2006; Rudner, 2002; Weiss, 1982). For this reason, question items are separately determined for each participant in adaptive tests. Thus, more efficient and accurate measurement is performed in adaptive tests in comparison to traditional tests (Cheng, Lin & Huang, 2009; Triantafillou, Georgiadou & Economides, 2008; Weiss, 1985).

The adaptive tests provide a more efficient measurement through very few items in comparison to conventional tests (Cheng, Lin & Huang, 2009; Koong & Wu, 2010; Kreitzber, Stocking & Swanson, 1978; Lunz, Bergstrom & Gershon, 1994; Weiss, 2004). In addition to this efficiency, measurement precision substantially increases as the questions providing most information about the ability level of a participant are asked. Moreover, the standard error rate of measurement falls thanks to the increase in efficiency ensured in the measurement. A decrease in standard error is an indicator that the ability levels of participants are measured more precisely (Bulut & Kan, 2012; Huang, Lin & Cheng, 2009; Liu & Chen, 2012; Vispoel, Rocklin & Wang, 1994; Reckase, 2010; Weiss, 1985).

Related Works

CAT systems are used in different fields such as education, health, certification, and undergraduate programs. Rios, Millán, Trella, Pérez-de-la-Cruz, and Conejo (1999) focused on the development of a test generation system module, one of the three components of the TREE (TRaining of European Environmental trainers and technicians in order to disseminate multinational skills between European countries) project. The test generation system was developed as a CAT system independent from the TREE system. The results showed that adaptive testing systems could be successfully integrated into e-learning systems, and more efficient evaluation could be ensured by this means. Eggen and Straetmans (2000) compared paper-and-pencil tests and a CAT application an exam, whereby students would be placed in courses based on their ability levels. According to the study, a decrease of 22% to 44% took place in the CAT application in comparison to paper-and-pencil tests in terms of the number of question items required. Gouli, Kornilakis, Papinakolaou, and Grigoriadou (2001) focused on the CAT module framework, which they integrated into the INSPIRE system that had been designed as an adaptive hypermedia teaching system. That study described process steps in the process of design and implementation of this module, which was integrated into the INSPIRE system. No evaluation was performed concerning that system. López-Cuadrado, Perez, Vadillo, and Arruabarrena (2002) focused on the development process of the CAT system, which they integrated into the adaptive hypermedia system developed for language teaching. No evaluation was performed concerning that system. Fliege, Becker, Walter, Bjorner, Klapp, and Rose (2005) developed a CAT application for the diagnosis of depression symptoms. According to the results of the simulation carried out in that study, ability level could be determined at a high reliability level with an average of 6 items. López-Cuadrado, Armendariz, and Perez (2006) focused on the architecture of a CAT system they named GenTAI and integration of it into an e-learning environment. This system was not evaluated in any study. Lilley (2007) conducted a CAT application and compared it with the classical test. In this study, a test consisting of 20 questions total was implemented. Ten of these questions were prepared in the form of an adaptive test, and the remaining 10 were prepared in the form of a classic test. A Likert-type scale was used for determining the views of students about these tests. The students were asked to compare the above mentioned two test systems in terms of the general difficulty levels of questions and tests. The research findings demonstrated that questions provided in the CAT section were more suitable for the ability levels of participants and participants had quite positive attitudes concerning the CAT application. Walter, Becker, Bjorner, Fliege, Klapp, and Rose (2007) developed a CAT application for anxiety measurement. The sample of the study included 2348 psychiatry and psychosomatic patients, where a total of 13 scales and 81 items were considered enough to measure anxiety. Simulation activity showed that ability level could be measured at high precision through 6 to 8 items. Then, the results of the anxiety-CAT application and those of the classical anxiety inventory were compared through 102 clinical patients. According to research results, anxiety-CAT application can be used for the distinction of patients with a mental health problem in a reliable manner, as can the classical inventories. Choi, Reise, Pilkonis, Hays, and Cella, D (2010) carried out a study where a fixed-length test and CAT application were compared within the scope of the development of an emotional depression scale. A depressive symptom scale composed of 28 items was evaluated within the scope of that study. According to the research results, there was quite a high relationship between all short forms and the results of CAT application. In that study the CAT application yielded better results in all fields in comparison to the short forms. Frey and Seitz (2011) focused on the usability of a multi-dimensional CAT system for evaluating the literacy of students within PISA (Programme for International Student Assessment). When the exam held through a classical testing system was compared with the results of a CAT exam, CAT application was seen to have increased the measurement efficiency by 74%. In addition, it was seen that the CAT application decreased the number of question items addressed to students from 56 to 22, without any loss in the measurement precision. Öztuna (2008) developed and assessed a CAT system to determine the level of disability in low back pain and osteoarthritis of the knee. The results of this study showed that the CAT system can determine the levels of disability by fewer items than a classical test, and there is higher degree of cohesiveness with CAT and classical test results. Taking into account the different ability estimation procedures and test termination rules, Kalender (2011), the CAT ability level estimates and paper-and-pencil format of the student selection examination science subtest were compared. The results of the study showed that the CAT system provides more reliable ability level estimates with less items when compared to paper-and-pencil test format.

Considering the literature, CAT systems are commonly studied abroad. However, this is quite a new study area for our country. The aim of this study is to contribute to the use of CAT systems in our country's curriculum. To this end, a CAT system that can be used in the secondary education of our country's curriculum was designed and assessed. In this sense, this study is an attempt to answer the following research question in order to evaluate the system called UZWEBMAT-CAT:

- What is the reliability level of the CAT system developed for the unit of probability covered in the 11th grade mathematics curriculum?

Method

Research Design

In this study, a CAT system aimed at the sub-topics (permutation, combination, binomial expansion, and probability) of the unit of probability covered in the 11th grade mathematics course was developed. The developed CAT system's measurement accuracy was evaluated for all ability levels from low to high. To that end, the exam application was conducted at a high school located in Trabzon, Turkey, in the fall semester of the 2011-2012 academic year. The reliability coefficient values of each student test were calculated using the data obtained from exam application.

Procedure

The development of the CAT system

Forming an item pool. One of the most important elements of a CAT system is forming an item pool. The item pool consists of items concerning the characteristic that is planned to be measured. For this study, the characteristics planned to be measured are content knowledge, which include permutation, combination, binomial expansion, and probability. In the absence of items of known parameters (difficulty, item discrimination, guessing) and measure content knowledge of these topics, the new items were created. Items should be implemented on a certain number sample group to determine the parameters. The sample must be educated on these topics to answer these items. Therefore, 11th and 12th grade students were selected as the sample group to determine item parameters. The created test sheets were implemented at 11 different high schools located in Trabzon, Turkey, during the fall semester of the 2010-2011 academic year. The number of students undergoing each item varied between 605 and 654. The total number of students participating in the activity was 3146.

Analyzing the question items by the item response theory. Adaptive tests use Item Response Theory (IRT) instead of classical test theory (Wise & Kingsbury, 2000). The reason for this is that IRT has models that do not use the statistics in the group like classical test theory. This situation enables individual assessment (Hambleton, Swaminathan & Rogers, 1991; Ríos et al., 1999; Ponsoda, 2000; Weiss, 2004; Marinagi, Kaburlasos & Tsoukalas, 2007). Item analysis was performed according to IRT after implementation of the created items at schools. IRT has two assumptions to fulfill: unidimensionality and local independence. Unidimensionality shows that items of the test measure only one character. Local independence refers to answers given to items of the test being statistically independent. The existence of a dominant factor over test answers is regarded as evidence of unidimensionality of the test. Also, fulfillment of the unidimensionality assumption shows that the local independence assumption is also fulfilled. (Hambleton et al., 1991). All test sheets developed in this study were subjected to factor analyses, and whether or not these tests were unidimensional was searched. The results of the factor analyses showed that all tests were unidimensional. In other words, each of the created tests (permutation, combination, binomial expansion, and probability) measured its own content knowledge. Thus, assumption of local independence was also fulfilled.

Investigating model fit. The IRT model, based on which the answers obtained from the sub-tests of the unit of probability would be analyzed, was determined through model fit statistics. The MULTILOG 7.03 program was used to determine the model fit as well as the item parameters. All test sheets were analyzed, and the IRT model to be employed was determined. Analysis results demonstrated that the tests were fit for the 3-Parameter Logistic (3PL) model.

Examining item parameters. The parameters of each item need to be examined in order to decide on the items to be used in an adaptive testing application. Some items may not be in the form of an "S", which is the ideal form of the item characteristic curve. The reason for this may be that the said items have low discrimination and low difficulty values, or have a high chance parameter. Apart from that, the items with a negative discrimination value are also problematic items (Önder, 2007). The items considered problematic are removed from the test. In this sense, the item characteristic curves and item parameters of the items included in the tests were examined. To this end, the MULTILOG 7.03 program was used. The item characteristic curves of a sample test randomly chosen among tests are given below. Figure 1 shows the item characteristic curves of all items of the Permutation-8 test.

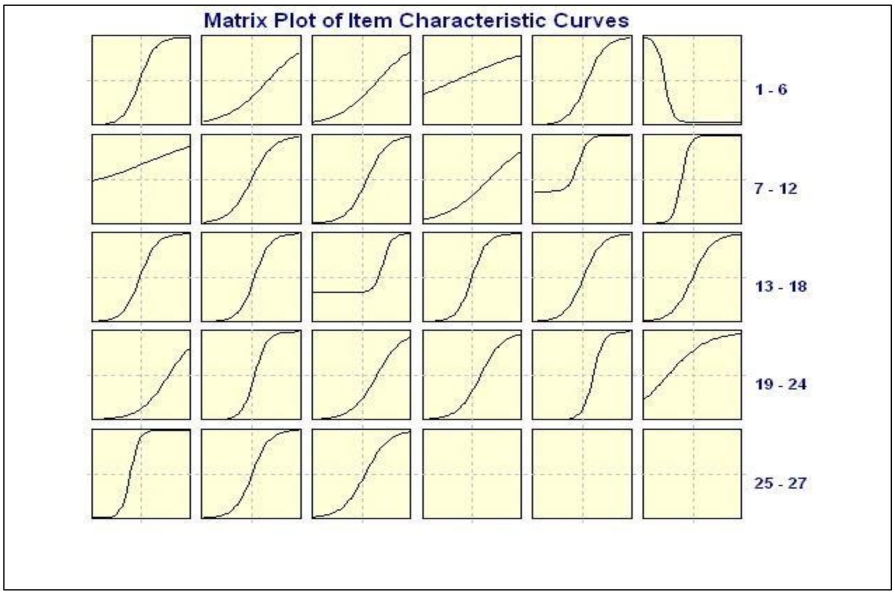

Figure 1. The item characteristic curves of all items of the permutation-8 test

As is seen in Figure 1, the item characteristic curve of the 6th item is a reverse S. This is because this item has a negative discrimination value. Thus, the 6th item was found to be problematic. Item parameters were examined after the item characteristic curves had been observed in order to better understand whether the items were problematic. Table 1 presents the item parameters of the Permutation-8 test based on 3PL.

Table 1

The Item Parameters of the Permutation-8 Test According to 3PL

Items	item discrimination (a)	difficulty (b)	guessing (c)
1.	1.187	-0.125	0
2.	0.459	0.966	0
3.	0.475	0.94	0
4.	0.193	-1.159	0
5.	1.039	0.254	0
6.	-2.36	-1.76	0
7.	0.285	0.235	0.379
8.	0.852	-0.025	0
9.	0.94	0.296	0
10.	0.438	1.029	0
11.	1.969	-0.282	0.362
12.	2.344	-0.726	0
13.	1.146	-0.106	0
14.	1.228	0.165	0
15.	2.125	1.3	0.328
16.	1.274	-0.023	0
17.	1.026	0.043	0
18.	0.952	-0.039	0
19.	0.659	1.69	0
20.	1.511	0.172	0
21.	0.765	0.906	0
22.	0.852	0.503	0

23.	1.802	0.645	0
24.	0.487	-1.585	0
25.	2.116	-0.689	0
26.	1.108	0.022	0
27.	0.821	0.203	0

The discrimination parameter of the 6th item was found to be lower than zero. Therefore, the 6th item was removed from the test. The remaining 26 items were included in the item pool. Such analyses were performed for all test sheets pertaining to the unit of probability. The analysis of test data showed that 86 items were problematic in all test sheets. These problematic items were removed from tests, and it was decided to include a total of 752 questions in the question pool. The distribution of questions in the final item pool by subtopics is as follows: permutation: 239; combination: 159; binominal expansion: 102; and probability: 252.

Designing and encoding the test system in the computer environment. This section focuses on the process steps in the process of designing and encoding a CAT system in the computer environment. The formation of the item pool was followed by the determination of the ability level estimation procedure, the item selection procedure, and the termination rule. In this study, Maximum Likelihood Estimation (MLE) was used as the ability level estimation procedure. Maximum Information Selection (MIS) was used in the item selection procedure. The fixed number termination rule was employed for terminating the testing session in the present study. Thus, the numbers of questions were as follows: permutation test: 15; combination test: 15; binominal expansion test: 15; probability test: 15; and end-of-unit test: 20.

Integrating the CAT system into UZWEBMAT. The developed CAT system was integrated to UZWEBMAT. UZWEBMAT is a Visual-Auditory-Kinesthetic (VAK) learning styles-based individualized, adaptive, and intelligent web-based mathematics learning environment (Özyurt, Özyurt & Baki, 2013). The UZWEBMAT consist of the subjects of permutation, combination, binomial expansion, and probability, which are covered in the unit of probability within the secondary education mathematics curriculum. The UZWEBMAT system is a dynamic learning environment that can adapt itself to the VAK learning styles and learning characteristics of students. The prepared CAT system was integrated into this environment as an assessment and measurement module. This module is named UZWEBMAT-CAT.

Sample

The developed testing system was employed in an exam in the school environment, and the obtained data were evaluated. The exam was conducted at a high school located in Trabzon, Turkey, in the fall semester of the 2011-2012 academic year. A total of 84 11th grade students took part in the exam. The names of students were kept confidential. Their names were encoded as Std1, Std2,..., Std84 in the present study. All actual information about the students was kept confidential in all figures.

Research Instruments

System records were used for data collection. System records contain level, score, exam session standard error value. It also includes learning style information belonging to each student in all exams.

Data Analysis

The reliability of IRT-based CATs is measured via standard error amount. As the standard error ratio decreases, the reliability of the test increases. Standard error is calculated as inversely proportional to the square root of the knowledge amount provided by the test (Hambleton, et al., 1991). Standard error amounts in IRT were converted to a classical reliability coefficient based on the formula suggested by Thissen (2000) in order for them to be better understood. Classical reliability coefficient ranges are as follows: 0-0.4: unreliable; 0.4-0.6: low reliability; 0.6-0.8 fair reliability; 0.8-1.0: high reliability (Özdamar, 2004). The relationship between standard error amount and reliability coefficient is as follows:

$$r = 1 - SE(\theta)^2 \quad (1)$$

r: Reliability coefficient

$SE(\theta)$: The standard error amount of the test at θ ability level

Results

The functioning and reliability of UZWEBMAT-CAT

This section provides UZWEBMAT-CAT's function and findings in relation to testing reliability. The functioning of a testing system - how the system approaches the real ability level over the given responses and the meantime change in reliability coefficient value - were presented to concretize the change. Exam data of three learners with different ability levels was used to show how UZWEBMAT-CAT moves from lower to higher levels. Figure 2 presents the data of three different students at the end- of-unit test. Std84 represented the higher level, while Std11 and Std80 represented intermediate and lower levels, respectively.

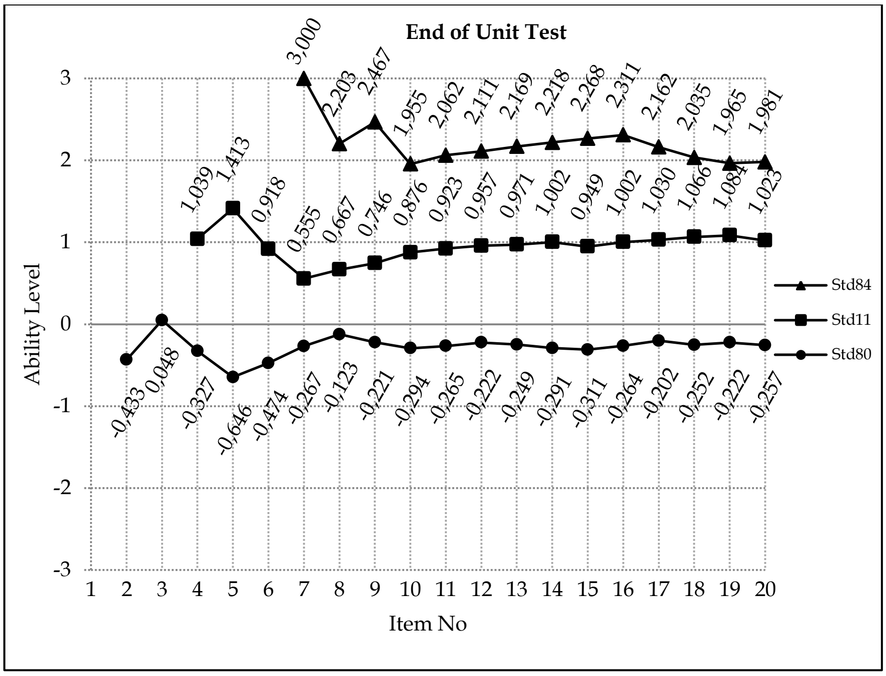

Figure 2. The comparison of the progress of Std84, Std11, and Std80 in the end-of-unit test

As seen in Figure 2, first ability estimations of Std84, Std11, and Std80 were made in the 7th, 4th, and 2nd questions, respectively. This is because the MLE procedure needs at least one correct and one incorrect response to estimate the level (Weiss, 1982). The first ability level estimation of Std84 was 3. According to the responses of Std84, UZWEBMAT-CAT either increased or decreased difficulty levels of questions and estimated the ability level in each step. Considering the general situation of Std84 in this exam and their ability level at the end of the exam (1981), it is possible to say that they had a high level of ability. According to the responses of Std11 and considering their general situation in the end-of-unit test and final level of ability (1,023), it is possible to say that they had a medium level of ability. Finally, considering the general situation of Std80 in the end-of-unit test and final level of ability (-0.257), it is possible to say that they had a low level of ability. Examining Figure 2 in detail, the extent of change in ability level estimations of these three learners with different levels is remarkable. As the test progresses, the difference between ability level estimations of each learner reduces gradually. This means that UZWEBMAT-CAT takes a step closer to the real ability levels of learners that were being estimated. Thus, the range in which real ability level is estimated shrinks. The estimation obtained at the end of the test is vitally close to the value of their real ability levels. This can be inferred from standard errors and reliability coefficient values calculated depending on this fact. The reliability of the system was investigated by using the system records obtained from the exam conducted through the UZWEBMAT-CAT. (The reliability of the system was investigated using system records obtained from exam implementation carried out via UZWEBMAT-CAT and detailed below).

For permutation, the reliability coefficient of the test taken by 3 students is between 0.78 and 0.79; the reliability coefficient of the test taken by 5 students is between 0.83 and 0.89; and the reliability coefficient of the test taken by remaining 76 students is between 0.90 and 0.96. Based on these values, it is seen that the test taken by 3 students for permutation is fairly reliable, while the tests taken by the other student have a very high reliability. For combination, the reliability coefficient of the test taken by 7 students is between 0.83 and 0.89; and the reliability coefficient of the test taken by remaining 77 students is between 0.90 and 0.97. Based on these values, the tests taken by all students for the combination test have a very high reliability. For binominal expansion the reliability coefficient of the test taken by 3 students is between 0.76 and 0.78; the reliability coefficient of the test taken by 49 students is between 0.81 and 0.89; and the reliability coefficient of the test taken by the other 32 students is between 0.90 and 0.95. Based on these values, it the tests taken by 3 students for binominal expansion tests are fairly reliable, and the tests taken by the remaining 81 students is very reliable. For probability, the reliability coefficient of the test taken by 1 student is 0.78; the reliability coefficient of the test taken by 16 students is between 0.80 and 0.89; and the reliability coefficient of the test taken by the remaining 67 students is between 0.90 and 0.97. Based on these values, the test taken by 1 student for binominal expansion is fairly reliable, and the tests taken by the remaining 83 students are very reliable. Finally, for end-of-unit, the reliability coefficient of the test taken by 2 students is between 0.76 and 0.78; the reliability coefficient of the test taken by 15 students is between 0.80 and 0.89; and the reliability coefficient of the test taken by the remaining 67 students is between 0.90 and 0.97. Based on these values, the test taken by 2 students for the end-of-unit test is fairly reliable, and the tests taken by the remaining 82 students are very reliable. Table 2 shows minimum, maximum, average standard error, and traditional standard coefficient values for each test between ability levels -3 and +3. In Table 2, "…" was used since there is no student estimated with an ability level between -3 and -1.

Table 2

Standard Error Amounts and Reliability Coefficients in All Tests

Ability Level			(-3,-2)	(-2,-1)	(-1,0)	(0,1)	(1,2)	(2,3)
Permutation test	Standard error amount	Min	…	…	0.21	0.21	0.21	0.24
		Max	…	…	0.25	0.25	0.47	0.45
		Average	…	…	0.238	0.237	0.235	0.32
	Traditional reliability coefficient	Min	…	…	0.94	0.93	0.78	0.79
		Max	…	…	0.96	0.96	0.96	0.94
		Average	…	…	0.943	0.943	0.944	0.897
Combination test	Standard error amount	Min	…	…	0.17	0.18	0.18	0.2
		Max	…	…	0.22	0.24	0.27	0.4
		Average	…	…	0.175	0.21	0.24	0.29
	Traditional reliability coefficient	Min	…	…	0.95	0.94	0.83	0.84
		Max	…	…	0.97	0.97	0.97	0.96
		Average	…	…	0.967	0.955	0.942	0.915
Binomial Expansion test	Standard error amount	Min	…	…	0.23	0.25	0.24	0.25
		Max	…	…	0.41	0.41	0.41	0.4
		Average	…	…	0.264	0.284	0.34	0.4
	Traditional reliability coefficient	Min	…	…	0.83	0.84	0.83	0.76
		Max	…	…	0.95	0.94	0.94	0.94
		Average	…	…	0.93	0.919	0.884	0.84
Probability test	Standard error amount	Min	…	…	0.17	0.17	0.17	0.19
		Max	…	…	0.24	0.24	0.41	0.47
		Average	…	…	0.187	0.191	0.253	0.359
	Traditional reliability coefficient	Min	…	…	0.94	0.94	0.83	0.78
		Max	…	…	0.97	0.97	0.97	0.96
		Average	…	…	0.964	0.963	0.935	0.87
End of Unit test	Standard error amount	Min	…	…	0.18	0.19	0.16	0.25
		Max	…	…	0.26	0.42	0.47	0.49
		Average	…	…	0.208	0.252	0.286	0.35
	Traditional reliability coefficient	Min	…	…	0.93	0.83	0.78	0.76
		Max	…	…	0.97	0.96	0.97	0.94
		Average	…	…	0.956	0.936	0.917	0.877

Discussion and Conclusion

The research findings show that the developed UZWEBMAT-CAT application generates highly reliable results. "Performing high precision and high reliability measurement through adaptive testing", which was suggested in the literature, was realized in the UZWEBMAT-CAT application (Eggen & Straetmans 2000; Choi et al., 2010; Frey & Seitz, 2011). The results of the UZWEBMAT-CAT application show parallelism to the results of the above-mentioned studies.

Weiss (1985) argues that CAT applications provide high measurement precision because they provide each individual with questions suitable for their own levels. The UZWEBMAT-CAT application also ensured high measurement efficiency because it provided every individual with questions suitable for their own ability levels. Another feature of CAT applications focused on in the literature is that they perform a very precise measurement at all ability levels, from low to high, because they provide every individual with tests suitable for their own ability levels (Eggen & Straetmans 2000; Frey & Seitz, 2011; Hambleton, 1990; Weiss, 1982; Wise & Kingsbury, 2000). The data obtained from the UZWEBMAT-CAT application support this conclusion. As a matter of fact, according to the data obtained from this application, ability level and reliability coefficient ranges of each test are as follows. Ability levels: between -0.618 and 2.437 for permutation test; between -0.427 and 2.481 for combination test; between -0.784 and 3 for binominal expansion test; between -0.71 and 2.436 for probability test; and between -0.401 and 2.312 for end-of-unit test. Reliability coefficients: between 0.78 and 0.96 for permutation test; between 0.83 and 0.97 for combination test; between 0.76 and 0.95 for binominal expansion test; between 0.78 and 0.97 for probability test; and between 0.76 and 0.97 for end-of-unit test. Based on all of these values, it can be said that subject tests perform high precision measurements for all ability levels. This study, in relation to the assessment of UZWEBMAT-CAT, prove that CAT systems can safely and efficiently be employed within Turkish curriculum.

References

Baki, A. (2008). *Teoriden pratiğe matematik eğitimi [Mathematics education from theory to practice]*. Ankara: Harf Eğitim Yayıncılık.

Bejar, I. I., Weiss, D. J., & Gialluca, K .A. (1977). An information comparison of conventional and adaptive tests in the measurement of classroom achievement (RR 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Choi, S. W., Reise, S. P., Pilkonis P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125-136.

Fliege, H., Becker, J., Walter, O. B., Bjorner J. B., Klapp, B. F., & Rose M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*(1), 2277–2291.

Frey, A., & Seitz, N .N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment. *Educational and Psychological Measurement, 71*(3), 503-522.

Georgiadou, E., Triantafillou, E., & Economides, A.A. (2006). Evaluation parameters for computer-adaptive testing, *British Journal of Educational Technology, 37*(2), 261-278.

Gouli, E., Kornilakis, H., Papinakolaou, K., & Grigoriadou, M. (2001). *Adaptive Assessment Improving Interaction in an Educational Hypermedia System*. Proceedings of the PanHellenic Conference with International Participation in Human-Computer Interaction, 217- 222.

Hambleton, R.K. (1990). Item response theory: Introduction and bibliography. *Psicothema, 2*(1), 97-107.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: SAGE Publications.

Huang, Y-M., Lin, Y-T., & Cheng, S-C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education, 52*(1), 53-67.

Koong C-S., & Wu, C-Y. (2010). An interactive item sharing website for creating and conducting on-line testing. *Computers & Education, 55*(2010), 131–144.

Lilley, M. (2007). *The development and application of computer-adaptive testing in a higher education environment*. Unpublished doctoral dissertation, Hertfordshire University, England.

Liu, Y-C., & Chen, P-J. (2012). Discovering discriminative test items for achievement tests. *Expert Systems with Applications, 39*(1), 1426-1434.

López-Cuadrado, J., Perez, T. A., Vadillo, J. A., & Arruabarrena, R. (2002*). Integrating adaptive testing in an educational system*. In Proc, First International Conference on Educational Technology in Cultural Context: ETCC2002, Joensuu, Finland, 133-149

López-Cuadrado, J., Armendariz, A., & Perez, T.A. (2006). Adaptive evaluation in an e-learning system architecture. *Current Developments in Technology-Assisted Education*, 1507-1511.

Lunz, M. E., Bergstrom, B. A., & Gershon, R. C. (1994). Computer adaptive testing. *International Journal of Educational Research, 21*(6), 623-634.

Marinagi, C. C., Kaburlasos, V. G., & Tsoukalas, V. T. (2007). *An architecture for an adaptive assessment tool*. 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee, WI.

Önder, İ. (2007). An investigation of goodness of model data fit. *Hacettepe University Journal of Education, 32*, 210-220.

Özdamar, K. (2004). *Paket programlarla veri analizi [Statistical data analysis through packages]*. Eskişehir: Kaan Yayıncılık.

Özyurt, Ö., Özyurt, H., & Baki, A. (2013). Design and development of an innovative individualized adaptive and intelligent e-learning system for teaching-learning of probability unit: Details of UZWEBMAT. *Expert Systems with Applications, 40*(8), 2914-2940.

Öztuna, D. (2008). *Kas iskelet sistem sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması [Implementing computer adaptive testing method to estimate disability levels in musculoskeletal system disorders]*. Unpublished doctoral dissertation, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü. Ankara.

Ponsoda, V. (2000). Overview of computerized adaptive testing special section. *Psicologica, 21*(1), 115-120.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Rios, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Conejo, R. (1999). *Internet based evaluation system*. In Proceedings of the 9th World Conference of Artificial Intelligence and Education AIED'99, 387-394.

Rudner, L.M. (2002). *An examination of decision-theory adaptive testing procedures*. Conference of American Educational Research Association, New Orleans, April 1-5.

Semerci, Ç., & Bektaş, C., 2005. New approaches in ensuring the validity of internet-based measurements. *The Turkish Online Journal of Educational Technology, 4*(1), 130-134.

Thissen, D. (2000). Reliability and measurement precision. In: H. Wainer (Ed). *Computerized adaptive testing: A primer (159-183)*, Mahwah, NJ: Lawrence Erlbaum Associates.

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, 7*(1), 53-79.

Walter O. B., Becker J., Bjorner J. B., Fliege, H, Klapp, B. F., & Rose M. (2007). Development and evaluation of a computer adaptive test for 'anxiety' (Anxiety-CAT), *Quality of Life Research, 16*(1), 143–155.

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

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Wise, S. L., & Kingsbury, G .G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135-155.

Öğrencilerin Olasılık Ünitesi Bilgi Seviyelerinin Bilgisayarlaştırılmış Uyarlanabilir Test ile Kestirilmesi

Atıf:

Özyurt, H. & Özyurt, Ö. (2015). Ability Level Estimation of Students on Probability Unit via Computerized Adaptive Testing. *Eurasian Journal of Educational Research*, *58*, 27-44. http://dx.doi.org/10.14689/ejer.2015.58.7

Özet

Problem Durumu: Öğrenme-öğretme faaliyetleri beraberinde bu faaliyetlerin amacına ulaşıp ulaşmadığının belirlenmesi ihtiyacını getirmektedir. Bu ihtiyaç ölçme ve değerlendirme faaliyetleri ile giderilmektedir. Bu nedenle ölçme ve değerlendirme eğitim-öğretim sürecinin ayrılmaz bir parçası haline gelmiştir. Eğitimde ölçme ve değerlendirme amacıyla çeşitli yöntemler kullanılmaktadır. Bu yöntemlerden biri de çoktan seçmeli testlerdir. Bu testler özellikle geniş ölçekli sınav uygulamaları için oldukça kullanışlıdır. Yüz yüze eğitimde olduğu gibi internet temelli eğitimde de tıpkı klasik eğitim yöntemlerinde olduğu gibi ölçmeye ihtiyaç duyulmaktadır. İnternet teknolojisinin getirdiği avantajlarla bilgisayar ortamında çok sayıda öğrencinin aynı anda sınav olması mümkündür. Ancak bu geleneksel ölçme-değerlendirme şekli, bireysel öğrenme özelliklerinin ön plan çıktığı günümüz modern eğitim anlayışına ters düşmektedir. Bu durum, geleneksel testlerin yeni eğitim anlayışına daha uygun olan uyarlanabilir testlere dönüşmesini hızlandırmıştır. Bu nedenle bireylerin bilgi seviyelerine göre soruların zorluk düzeyini belirleyen Bilgisayarlaştırılmış Uyarlanabilir Test (BUT) sistemleri giderek yaygınlaşmaktadır. BUT'larda testin zorluğu, katılımcının bilgi seviyesine göre dinamik olarak değişir. Katılımcı soruyu doğru yanıtlarsa sonraki soru daha zor bir olacak şekilde seçilir. Katılımcı soruyu yanlış yanıtlarsa sonraki soru daha kolay bir

soru olacak şekilde seçilir. Sorular seçilirken katılımcının bilgi seviyesi hakkında en fazla bilgiyi sağlayan maddeler seçilir. Böylece uyarlanabilir testlerle geleneksel testlere göre daha verimli ve hassas ölçüm yapılmaktadır. BUT sistemleri Madde Tepki Kuramı'nı (MTK) temel alan uygulamalardır. Yurt dışında oldukça yaygın olan BUT sistemleri Türkiye'de henüz yaygınlaşmamıştır. Bu nedenle Türk eğitim sistemi içerisinde kullanılabilecek bir BUT sisteminin geliştirilmesi ve değerlendirilmesi önem taşımaktadır.

Araştırmanın Amacı: Bu çalışmanın ana amacı Türk eğitim sisteminin alt basamaklarında kullanılabilecek (lise düzeyi) bir BUT sistemi geliştirmek ve güvenirlik düzeyi bakımından değerlendirmektir.

Yöntem: Bu çalışmada 11. sınıf matematik dersi olasılık ünitesinin alt konularına (permütasyon, kombinasyon, binom açılımı ve olasılık) yönelik bir BUT sistemi geliştirilmiştir. BUT sisteminin geliştirilmesi, soru maddelerinin geliştirilmesi, uygulanması ve MTK'ya göre analiz edilmesi gibi işlemlerden oluşan madde havuzu geliştirme aşaması, sistemin bilgisayar ortamında kodlanması ve uyarlanabilir zeki web tabanlı matematik öğrenme ortamı UZWEBMAT'a entegre edilmesi aşamalarından oluşmaktadır. Geliştirme aşamasından sonra BUT sisteminin düşükten yükseğe bütün bilgi seviyelerinde ölçüm hassasiyeti değerlendirilmiştir. Bunun için 2011-2012 öğretim yılı güz döneminde Türkiye'de Trabzon ilindeki bir lisede sınav uygulaması yapılmıştır. Sınav uygulamasına toplam 84 11. sınıf öğrencisi katılmıştır. Sınav uygulamasından elde edilen veriler kullanılarak her bir öğrencinin aldığı testlerin güvenirlik katsayıları hesaplanmıştır. MTK temelli BUT'ların güvenirliği, standart hata miktarı ile ölçülmektedir. Standart hata oranı düştükçe testin güvenirliği artmaktadır. Standart hata miktarı klasik güvenirlik katsayısına dönüştürülerek sunulmuştur. Veri toplama aracı olarak sistem kayıtları kullanılmıştır.

Araştırmanın Bulguları: Test sisteminin işleyişi, sistemin verilen yanıtlara göre her adımda gerçek bilgi seviyesine nasıl yaklaştığını ve bunu yaparken güvenirlik katsayısındaki değişimi somutlaştırmak için sunulmuştur. UZWEBMAT-CAT'in düşükten yükseğe bütün bilgi seviyelerinde nasıl hareket ettiğini göstermek için bilgi seviyeleri farklı üç öğrencinin ünite sonu testi verileri kullanılmıştır. Öğrenciler seçilirken yüksek, orta ve düşük bilgi seviyesi olarak nitelendirilebilecek öğrenciler olmasına dikkat edilmiştir. Farklı bilgi seviyesine sahip bu üç öğrencinin ilk sorudan son soruya doğru bilgi seviyesi kestirimlerindeki değişim miktarı dikkat çekmektedir. Test ilerledikçe her bir öğrencinin kendi bilgi seviyesi kestirimleri arasındaki fark giderek azalmaktadır. Bunun anlamı öğrencilerin kestirilmeye çalışılan gerçek bilgi seviyelerine UZWEBMAT-CAT'in her adımda biraz daha yaklaşmasıdır. Böylece gerçek bilgi seviyesinin kestirilmeye çalışıldığı aralık daralmaktadır. Testin sonunda elde edilen kestirim öğrencilerin her birinin gerçek bilgi seviyelerine çok çok yakın bir değer olmaktadır. Bu durum testin standart hata ve buna bağlı olarak hesaplanan güvenirlik katsayısı incelenerek anlaşılabilir. Permütasyon, kombinasyon, binom açılımı, olasılık ve ünite sonu testi olmak üzere toplam 5 ayrı sınav için öğrencilerin her birinin aldığı testin güvenirlik katsayıları hesaplanmıştır. Bunun için her bir testin standart hata miktarı klasik güvenirlik

katsayısına dönüştürülmüştür. Çalışmanın bulguları geliştirilen BUT sisteminin tüm konular için yüksek güvenirlikte testler ürettiğini göstermektedir. Her bir konuya ilişkin testlerin ortalama güvenirlik katsayıları 0.93, 0.93, 0.88, 0.93 ve 0.91 olarak hesaplanmıştır. Geliştirilen BUT sistemi düşükten yükseğe tüm bilgi seviyelerinde hassas ve güvenilir testler ürettiği görülmüştür.

Tartışma ve Sonuç: Literatürde BUT ile ilgili üzerinde durulan önemli özelliklerinden biri de her bir bireye kendi bilgi seviyesine uygun bir test sunduğu için düşükten yükseğe bütün bilgi sevilerinde çok hassas ölçüm yapma özelliğidir. Bu özelliğin geliştirilen BUT sistemi tarafından sağlandığı görülmüştür. Nitekim farklı bilgi seviyesine sahip öğrencilerin aldıkları testlerin yüksek güvenirliğe sahip olduğu görülmüştür. Literatürde üzerinde durulan bir diğer durum da BUT'ların her bireye kendi bilgi seviyesinde sorular yönelttiği için ölçümün standart hata miktarının düştüğü ve dolayısıyla ölçüm hassasiyetinin artmasıdır. Geliştirilen BUT sisteminin de her bir öğrenciye kendi bilgi seviyesinde sorular yönelttiği böylece ölçüm hassasiyetinin arttığı görülmüştür. Yapılan testlerin güvenirlik katsayılarının değer aralıklarına bakıldığında az sayıda oldukça güvenilir ve çok sayıda çok güvenilir aralıklarına giren değerler olduğu görülmektedir. UZWEBMAT-CAT'in değerlendirilmesiyle ilgili bu çalışma ile Türk eğitim sisteminde de BUT sistemlerinin güvenilir bir şekilde kullanılabileceği ortaya konulmuştur.

Anahtar Sözcükler: Bilgisayarlaştırılmış uyarlanabilir test, Bireysel değerlendirme, Bireysel farklılıklar, Bilgi seviyesi kestirimi, Uyarlanabilir test.