Journal of Teacher Education and Lifelong Learning (TELL)

Volume:5 Issue: 2 Year: 2023 Research Article ISSN: 2687-5713

Evaluating knowledge levels of students with a Computerized Adaptive Test[1]

Mustafa Yağcı [1]

[1] Kırşehir Ahi Evran University, Faculty of Engineering and Architecture, Kırşehir, Turkey
mustafayagci@ahievran.edu.tr

Article Info

Article History
Received:10/04/2023
Accepted: 12/12/2023
Published: 16/12/2023

Keywords:
Computerized adaptive test, Individualized testing, Item response theory, Paper-and-pencil test, Number of test items, Intelligent tutoring systems

ABSTRACT

The present study aims to develop and evaluate a Computerized Adaptive Test (CAT) system for multiple-choice success tests. In this regard the measurement results obtained from the CAT system based on Item Response Theory (IRT) and the test mode based on the Classical Test Theory (CTT) were compared in terms of the knowledge level of the students, reliability of measurement and number of items. The study was conducted on 873 students from a state-university in Turkey. The research took three years and involved three phases. According to the findings of the present study, with the developed CAT system, academic success levels of the students were determined with very high reliability. When all the items in the test were scored according to both constructs, the relationship between the scores obtained was found to be high and meaningful in a positive direction. Moreover, there is a high, positive, significant relationship between students' levels of knowledge that are estimated in CAT systems and the score of an achievement test they received from a Paper & Pencil test. Finally, the number of questions the students were exposed to in the CAT system was reduced by 50% compared to the test based on the CTT.

Citation: Yağcı, M. (2023). Evaluating knowledge levels of students with a Computerized Adaptive Test. *Journal of Teacher Education and Lifelong Learning*, *5*(2), 921-932.

[1] This study was supported by Kırşehir Ahi Evran University BAP coordinatorship with project number MMF.A4.18.008.

1. INTRODUCTION

The idea of the tailored test first emerged in Binet's work to develop the IQ test in 1905 (Binet and Simon, 1905). This test Binet prepared had an item pool prepared in advance, items classified based on the level of difficulty, starting options, a scoring method defined in advance, and a rule for the selection from a pool of the items to be asked and for a termination defined in advance. Even if this first application revealed about 100 years ago appears basic, it can be said that it created a foundation for applications today. However, Binet's tailored test did not arouse much interest in that period because of the questions in practice. Because an IQ estimation needs to be done accurately and quickly based on the responses an individual provides to a specific item and then a new item needs to be selected from the item pool and answered. This is why such an application was adopted in the 1970s with advancements in computer technology.

Considering the history of measuring, it is seen that tests were developed in the context of two fundamental theories of testing, referred to as the Classical Test Theory (CTT) and the Item Response Theory (IRT) (Crocker and Algina, 1986). The score based on the CTT in the measurement of achievement is found with the compilation of the scores that students received from the items. The test and item statistics were conducted over the total scores and item scores obtained in this manner. Therefore, the scores individuals received based on the CTT vary based on the level of difficulty of that test (Lord and Novick, 1968). The difficulty indexes of items, however, are generally not taken into account, and if items are not weighted when scored, the contribution of each item to the total score is the same amount. Because individuals at different ability or achievement levels are tested in these tests, numerous items that encompass a wide interval must be administered. For example, on a test the contains some very easy items because the probability of individuals with a high ability level of answering these items correctly is high, these items do not serve a purpose in distinguishing upper-level individuals from one another.

The IRT is a theory that specifies that there is a relationship between the proficiency levels of individuals measured with the test and the response behaviour to any of the items on the test, and it explains this relationship based on a probability model (Embretson and Reise, 2000; Hambleton, et al., 1991; Wainer et al., 1990). The contribution of each item to the score calculated based on the IRT is not the same; scores for items that are affected by the difficulty of items based on the model used, by the strength of item discrimination, and by the probability of responding correctly by guessing are in question. The traits of the invariance of the item parameters and the invariance of the ability parameter in the IRT are the most important scores that separate this theory from the classical test theory and that provide for its use in Computerized Adaptive Test (CAT) systems.

The invariance of the item parameters means the parameters used to define the items (item discrimination, difficulty index, etc.) can be independently estimated from the ability levels of individuals. Even if the items are administered only to a specific group, the item parameters can be estimated to encompass all ability levels with a nonlinear regression method (Kalender, 2011). The invariance of the ability parameter means the ability estimations for individuals can be made independently of the items asked individuals (Kalender, 2011). Even if each individual receives different items, the ability estimations are found on the same scale for each individual and thus are comparable. Even if there are many different IRT models, the models most frequently used and most appropriate for the tests coded as true/false are 1, 2, 3PL: One, Two, Three parameter Logistic Models. These models try to explain the relationship between the ability level of an individual and the probability of correctly answering this question, by using different parameters. These parameters are referred to as item discrimination (a), difficulty index (b), and guessing parameter (c). The guessing parameter is the probability for individuals in the lowest ability level to incidentally answer an item correctly. The 1PL model uses only the item difficulty index, the 2PL model takes the item discrimination level of the item into account along with the difficulty index, and the 3PL model considers guessing parameter in addition to these two parameters. The items used in the administering of CAT are previously administered to a group, and its parameters are included in the item pool in a defined state (calibrated) based on one of the models above.

An item characteristic function is defined for each of the items in these models. This function forms a relationship between the ability level of individuals, using the item parameters mentioned above and the possibility of correctly answering the item; and it is used for the ability estimation of individuals during the administering of CAT.

The item information function is used for the selection of items in the administering of CAT. The standard error of the estimation of ability is inversely proportional to the test information function. Based on this, using items that have a high information value increases the reliability of the ability estimation by decreasing standard error and ensures the conclusion of the test in a shorter amount of time.

1.1. Literature

Researches examining the applicability of tests designed based on the IRT and whether these tests are an alternative tool of measurement to tests prepared based on the CTT are as follows.

Pelanek (2016) designed a system of rating that dynamically matched the relationship between the ability level of the student and the question. As a result of applying this, in educational applications the Elo rating system is simple, substantial and efficient and therefore suitable for development. Cisar, Cisar and Pinter (2016) conducted an experimental study with two groups of students. The result of the research shows that the students who studied on computer adaptive testing are more successful than the students studying in the paper pencil test (PPT). Çelen and Aybek (2013) compared by individually scoring a test that measured the success in the measurement evaluation class based on the CTT and IRT methods. The BILOG-MG 3 software was used for analysis. As a result, a positive and meaningful relationship was found between the acquired scores. Huang et al. (2009) developed and implemented an adaptive testing system. The result of the research shows that the system can estimate students' abilities reliably and validly and conduct an adaptive test efficiently. With his CAT system, Tseng (2016) was able to classify students' vocabulary with far fewer questions. Özbaşı and Demirtaşlı (2015) tested the applicability of a computer literacy test as a CAT. The data were gathered with the SimulCAT simulation program and with the CAT system that Kalender (2011) designed. The average reliability level acquired from the administering the CAT was found to be significantly higher than the value obtained from the PPT. Chen and Chung (2008) developed a mobile English vocabulary learning system based on IRT. The result of the experimental research shows that the system positively affected the students' learning performance and learning interest in English vocabulary learning. Kuo et al., (2015) developed multidimensional CAT for Indonesia junior high school Biology. The study was carried out in two phases (simulation study and real data). The results show that the bias and standard error of the Biology-MCAT system are acceptable. Özyurt (2013) evaluated the level-setting of the designed adaptable test system for the knowledge levels of the students. It was found that as a result of the conducted application, the developed test system was able to very reliably level-set the knowledge levels of the students. The findings of the study of Petscher, Foorman and Truckenmiller (2017) indicated that examinees in the experimental condition took fewer passages and had more reliable estimates of their reading comprehension ability. Zhang and VanLehn (2017) found that students using adaptive question selection had reliably larger learning gains than students who received questions in a maladaptive order. Wu et al. (2018) argue that the computerized dynamic adaptive test with individualized prompts outperformed the other two instruction methods (the individualized instruction without adaptive dynamic assessment, and the traditional classroom remedial instruction).

It is understood from the literature that the CAT system can calculate the knowledge levels of students much quicker and more reliably. In addition to this, it is seen from research results and the theoretic framework that CAT systems are quite advantageous. Although there has recently been an increase in the number of studies conducted in this field, they are quite limited. The reason for this could be that the design of CAT requires specialization both on the topics of programming and measurement evaluation. When it is as such, researchers prefer test statistics with simulation software like SimulCAT rather than real, experimental studies with CAT. On the other hand, the formation of a item pool with real experimental applications requires a period about one or two years. Additionally, there is no platform on which researchers and

individuals who want to conduct this type of application can easily make designs.

The CAT system designed on the “Concerto” platform and the Computer Based Assessment (CBA) designed on the My-Sql platform by the researcher were used as the test tool in this study. In this context, the examination of the applicability and usability of CAT in the measurement of the knowledge levels of students will provide significant contributions to the field.

This study examined the applicability in a CAT system developed based on the IRT of the achievement test of the “Office Programs” course, a basic, required course taught at universities. All students must answer the entirety of the questions in these tests, generally administered as multiple choice, regardless of the level of proficiency. In other words, students are forced to answer more questions than necessary that fall above or below their proficiencies. This situation weakens the reliability of tests and the motivation of students in exams. To eliminate these limitations, testing the applicability as a CAT of the achievement test administered as a PPT constitutes the fundamental problem of this research.

1.2. Purpose of the Research

The aim of this study is to examine student achievement, which is estimated based on IRT in the CAT system and calculated based on CTT, which is one of the two equivalent measurement tools aimed at measuring the academic success of office programs. The initial hypothesis was stated as follows:

1. The reliability coefficient of the "office programs" achievement test carried out with the CAT system is higher than that of the KTK.

2. There is a positive and highly significant relationship between the scores calculated based on the IRT and CTT.

3. The average number of questions that students have to answer in the CAT system is less than the number of questions they answer according to the KTK.

2. METHOD

In this section, the research pattern, study group and teaching materials used in the study are explained. Data collection tools and statistical analyses of collected data are described.

2.2. Study Group

The research was carried out in three phases. 1st phase was conducted on 745 students receiving the “Office Programs” course and who were attending the Faculty of Education and Faculty of Arts and Sciences at a state university during of the 2018-2019 academic year. 2nd phase was conducted on 128 students receiving the “Office Programs” course and who were attending the Faculty of Education and the Faculty of Arts and Sciences at a state university during of the 2019-2020 academic year. The final phase was also conducted with 128 students in the 2nd phase.

2.3. Data Collection Process and Experimental Operation

1st phase: In this phase, first of all, an online exam system with multimedia support and a user-friendly interface was developed by the researcher.

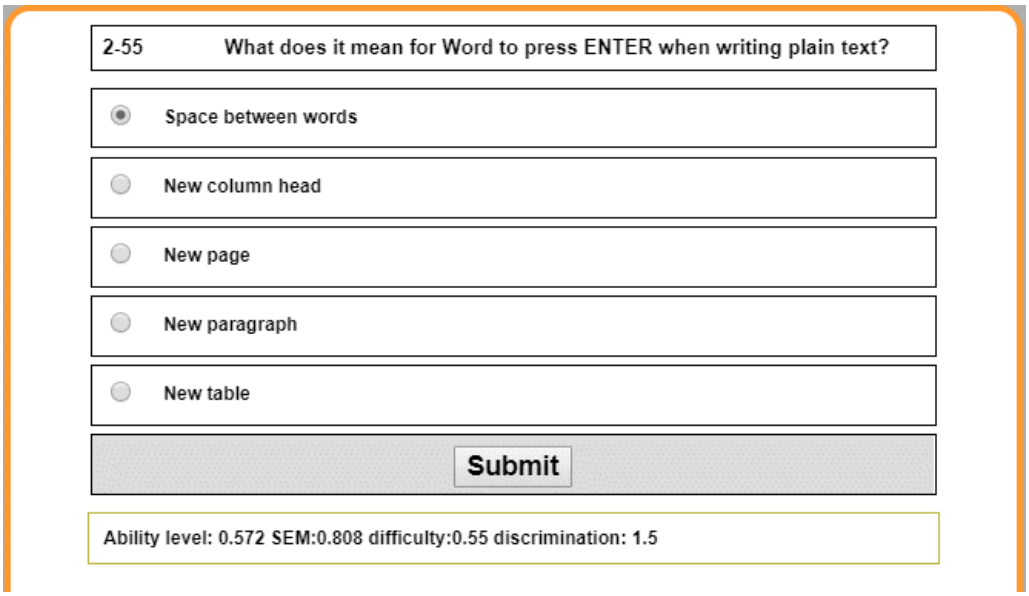
Then, the achievement test, consisting of 118 questions (a multiple choice test), was applied to a total of 745 students who took the "Office Programs" course at a state university. In the tests carried out throughout a 2-year process in 4 sessions, the item parameters were automatically calculated by the system based on the answers the students provided to the questions. Items whose difficulty and index of distinctiveness values were not suitable were again automatically removed from the question bank.

At the end of the conducted application, 43 items with index of distinctiveness index below 0.30 and difficulty index less than 0.25 greater than 0.9 were removed from the item pool. At the end of the first stage, an item pool of 75 questions calibrated with the 2-Parameter model was obtained. Because the 2PL model takes the item discrimination level of the item into account along with the difficulty index.

2nd phase: In this phase, first of all, the researcher designed an IRT-based CAT system. This system was designed on the open-sourced Concerto platform, developed by the University of Cambridge Psychometrics Centre. While the test codes were written with the "R" language on the Concerto platform, the tables, which are the cornerstones of the database, can be configured in a graphic interface just as in the My-Sql platform. Interface design can be done with HTML codes or graphic-based.

Rules of termination are generally chosen as fixed length and varying length in the body of literature. Standard error, most preferred in varying length conclusion rules, emerged as the conclusion rule. Additionally, it is recommended that the varying length conclusion rule be limited with a maximum number of items. This is why a standard error of 0.30 and a maximum of 40 items were specified in this study as the conclusion rule. The CAT module with the aforementioned traits was added to the online examination system designed in the first phase. Figure 1 is a screenshot from the CAT module.

Fig. 1. CAT screenshot



Then, CAT was administered with 128 students who were taking the Office Programs course in 6 different departments in the Faculty of Education and the Faculty of Arts and Sciences at a State University. In this administering, 75 questions that had been previously calibrated following the 2PL model were used. The ability estimation method has been specified as the "Maximum Likelihood", the question-selection method as the "Maximum Information", and the test termination rule as the "standard error falling below 0,30" or the number of administered questions reaching 40.

3rd phase: Forty questions that had homogeneous distribution were identified based on the difficulty (p) and item discrimination (r) levels of the 75 questions whose suitability is specified for the 2PL model. These 40 questions were administered again as a PPT 1 week after the CAT was administered to the student group in the 2nd phase.

3. FINDINGS

3.1 Findings for the First Sub-Purpose

Table 1 presents the knowledge level estimations, test scores, standard errors, and reliability coefficients of all students in the administering of the office programs test.

Table 1. *The knowledge level estimations, test scores, and standard error and reliability coefficient values for all students who took the test.*

Std.	N. of item	Theta	Std. error	Success grade (100)	Reliability coefficients	St.	N. of item	Theta	Std. error	Success grade (100)	Reliability coefficients

Std1	19	0.73	0.30	70	0.913	Std65	19	0.65	0.30	70	0.91
Std2	21	0.04	0.30	20	0.912	Std66	19	0.24	0.30	35	0.91
Std3	19	0.38	0.30	65	0.913	Std67	19	0.89	0.30	80	0.91
Std4	21	1.09	0.29	90	0.915	Std68	19	0.79	0.30	85	0.91
Std5	19	0.86	0.30	85	0.912	Std69	19	0.38	0.30	45	0.91
Std6	19	0.24	0.30	40	0.911	Std70	19	0.53	0.30	70	0.91
Std7	19	0.51	0.29	65	0.914	Std71	19	0.38	0.30	45	0.91
Std8	19	0.61	0.30	80	0.912	Std72	19	0.65	0.30	75	0.91
Std9	19	0.66	0.30	75	0.912	Std73	19	0.51	0.29	55	0.91
Std10	19	0.53	0.30	75	0.913	Std74	19	0.24	0.30	35	0.91
Std11	20	0.42	0.29	65	0.915	Std75	19	0.28	0.30	40	0.91
Std12	19	0.71	0.30	75	0.913	Std76	19	0.24	0.30	45	0.91
Std13	19	0.64	0.30	70	0.912	Std77	20	0.20	0.29	30	0.91
Std14	19	0.39	0.30	55	0.913	Std78	19	0.51	0.29	70	0.91
Std15	19	0.52	0.30	65	0.913	Std79	19	0.51	0.29	80	0.91
Std16	19	0.85	0.30	75	0.912	Std80	19	0.51	0.29	60	0.91
Std17	19	0.84	0.30	75	0.912	Std81	19	0.38	0.30	40	0.91
Std18	19	0.89	0.30	80	0.912	Std82	21	0.04	0.30	10	0.91
Std19	19	0.57	0.30	65	0.912	Std83	22	0.00	0.29	5	0.91
Std20	19	0.64	0.30	80	0.912	Std84	19	0.39	0.30	45	0.91
Std21	19	0.51	0.29	75	0.914	Std85	25	-0.21	0.30	5	0.91
Std22	19	0.71	0.30	85	0.912	Std86	19	0.24	0.30	30	0.91
Std23	19	0.59	0.30	75	0.912	Std87	25	-0.21	0.30	5	0.91
Std24	19	0.88	0.30	95	0.912	Std88	19	0.24	0.30	25	0.91
Std25	19	0.68	0.29	75	0.914	Std89	19	0.24	0.30	25	0.91
Std26	19	0.27	0.30	35	0.911	Std90	24	-0.05	0.29	5	0.92
Std27	19	0.65	0.30	70	0.912	Std91	19	0.38	0.30	45	0.91
Std28	20	0.46	0.29	60	0.915	Std92	19	0.68	0.30	65	0.91
Std29	19	0.64	0.29	70	0.914	Std93	21	0.16	0.29	60	0.92
Std30	19	0.63	0.30	75	0.912	Std94	19	0.38	0.30	55	0.91
Std31	21	0.19	0.29	25	0.915	Std95	19	0.76	0.30	50	0.91
Std32	19	0.25	0.30	30	0.911	Std96	19	0.24	0.30	25	0.91
Std33	19	0.51	0.29	50	0.914	Std97	19	0.40	0.30	70	0.91
Std34	19	0.51	0.29	65	0.914	Std98	21	0.04	0.30	10	0.91
Std35	20	0.55	0.29	70	0.915	Std99	19	0.51	0.29	75	0.91
Std36	19	0.82	0.30	90	0.912	Std100	21	0.04	0.30	10	0.91
Std37	20	0.77	0.29	85	0.915	Std101	19	0.24	0.30	25	0.91
Std38	19	0.58	0.30	60	0.912	Std102	19	0.38	0.30	60	0.91
Std39	19	0.65	0.29	65	0.914	Std103	23	-0.02	0.29	5	0.92
Std40	19	0.38	0.30	60	0.913	Std104	21	0.04	0.30	5	0.91
Std41	27	-0.25	0.29	15	0.914	Std105	19	0.38	0.30	60	0.91
Std42	19	0.39	0.30	45	0.913	Std106	20	0.25	0.30	50	0.91
Std43	20	0.20	0.29	30	0.914	Std107	21	0.06	0.30	10	0.91
Std44	19	0.51	0.29	55	0.914	Std108	30	-0.43	0.31	5	0.91
Std45	21	0.04	0.30	10	0.912	Std109	19	0.89	0.30	75	0.91
Std46	19	0.56	0.29	65	0.914	Std110	19	0.38	0.30	50	0.91
Std47	19	0.83	0.30	70	0.912	Std111	19	0.38	0.30	60	0.91
Std48	19	0.39	0.30	30	0.913	Std112	19	0.65	0.30	60	0.91

Std49	19	0.65	0.30	75	0.912	Std113	20	0.75	0.29	65	0.92
Std50	21	0.16	0.29	20	0.916	Std114	21	0.04	0.30	5	0.91
Std51	19	0.51	0.29	65	0.914	Std115	21	0.16	0.29	25	0.92
Std52	19	0.65	0.30	55	0.912	Std116	30	-0.30	0.29	15	0.92
Std53	20	0.45	0.29	75	0.915	Std117	19	0.51	0.29	80	0.91
Std54	19	0.86	0.30	85	0.912	Std118	25	-0.20	0.30	70	0.91
Std55	19	0.51	0.29	60	0.914	Std119	30	-0.43	0.31	60	0.91
Std56	19	0.50	0.30	65	0.912	Std120	23	-0.03	0.29	60	0.92
Std57	30	0.59	0.33	5	0.893	Std121	19	0.24	0.30	65	0.91
Std58	19	0.53	0.29	65	0.914	Std122	19	0.38	0.30	80	0.91
Std59	19	0.51	0.29	62	0.914	Std123	19	0.43	0.30	70	0.91
Std60	19	0.50	0.30	70	0.912	Std124	19	0.24	0.30	60	0.91
Std61	20	0.20	0.29	25	0.914	Std125	20	0.20	0.29	65	0.91
Std62	21	0.04	0.30	10	0.912	Std126	19	0.51	0.29	75	0.91
Std63	19	0.51	0.29	70	0.914	Std127	21	0.16	0.29	65	0.92
Std64	20	0.20	0.29	30	0.914	Std128	21	0.04	0.30	45	0.91

It is seen from Table 1 that the reliability coefficient of the achievement test received values between 0.893 and 0.917. The arithmetical average of the reliability coefficient of the test is 0.913. According to these values, it can be said that the developed CAT system measured the knowledge levels of the students to a highly reliable degree.

Table 2 presents the progress and knowledge level estimations in the CAT system of Std105, the parameters of each question, the given responses, the amount of standard error, and the reliability coefficient provisions of this standard error.

Table 2. *Details of the administering of the end of the unit exam for Std105.*

Number	Item number	p	True/False	Theta	Std. error	Reliability coefficients
1	59	0.25	0	-------	--------	
2	42	0.3	1	-0.41	0.83	0.31
3	33	0.35	0	0.13	0.69	0.53
4	58	0.35	1	-0.12	0.63	0.61
5	41	0.4	1	0.20	0.56	0.69
6	62	0.4	0	0.42	0.51	0.74
7	74	0.45	1	0.24	0.48	0.77
8	44	0.55	1	0.41	0.45	0.80
9	53	0.45	0	0.56	0.43	0.81
10	55	0.55	1	0.43	0.41	0.83
11	60	0.65	1	0.55	0.39	0.85
12	12	0.6	0	0.65	0.38	0.86
13	14	0.5	0	0.55	0.36	0.87
14	49	0.6	1	0.47	0.35	0.88
15	7	0.5	0	0.55	0.34	0.89
16	45	0.65	0	0.48	0.33	0.89
17	57	0.65	0	0.40	0.32	0.90
18	48	0.7	0	0.34	0.31	0.90
19	19	0.75	0	0.43	0.30	0.91
20	51	0.7	0	0.38	0.30	0.91

According to Table 2, one of the items with the lowest difficulty index was asked to the individual as

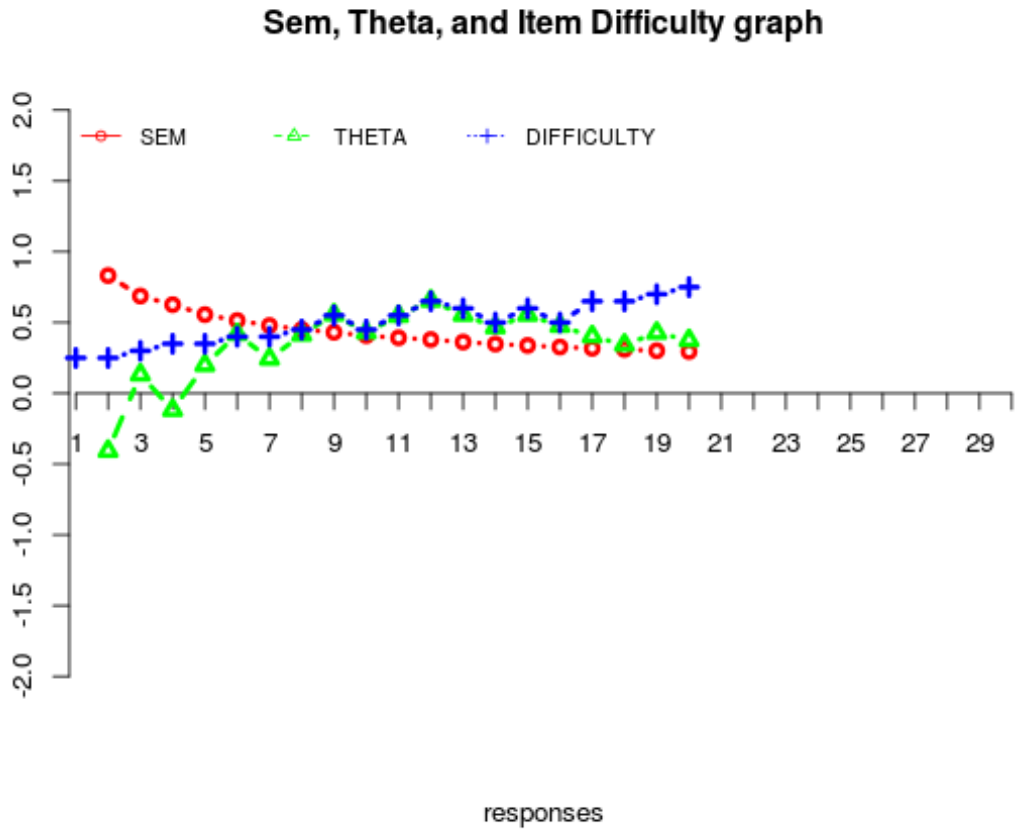
the first item (p = 0.25). Upon an individual incorrectly answering this item, the ability is attempted to be estimated by asking an item on the p=0.3 level and by providing a true/false condition, which is a condition for the maximum likelihood method (MLE). Upon the provision of a correct answer to the 2nd question by the individual, the first ability estimation was acquired as -0.407 (*SE*=0.832). The difficulty index of a subsequent item chosen based on the item information function was 0.35. The reason that the ability estimations of individuals and the difficulty indexes of the items do not completely overlap here is that in the phase of the selection of a subsequent item, not just difficulty index but also item discrimination are taken into consideration.

As can be seen in Table 2, 20 items were administered to the students in the test, and the test was concluded with the standard error dropping below 0.30 and the ability level of the individual being estimated as 0.375. It is seen that the difference between the knowledge level estimation values decreases towards the end of the test. Accordingly, after each step of the test, results even closer to the real knowledge level were obtained.

In addition, it is seen that the value of standard error decreased from the first step of the test towards the last and that the reliability coefficient value increased in connection with this. In the last step of the test, the standard error value decreased to 0.295, while the reliability coefficient increased to 0.913. As a result, it can be said that the test is very reliable and therefore the last estimated knowledge level value (0.375) is the closest value to the real knowledge level of Std105.

Figure 2 visually presents the ability, estimations of standard error, and change in difficulty index of the questions for the individuals whose test advancement report is provided in Table 2.

Figure 2. Standard error, ability and item difficulty index progress



According to Chart 1, the range of ability estimation narrows as the test progresses. While the difficulty level difference between the items in the first questions is larger, it gets smaller as the test progresses, and the individual's ability estimation becomes more and more precise.

3.2. Findings for the Second Sub-Purpose

The relationship between the T scores obtained from the CAT based on the IRT and the paper pencil test based on the CTK was examined with the Spearman Rank Differences Correlation Coefficient and the results are given in Table 3.

Table 3. *Spearman rank-order correlation coefficient between ability scores estimated according to the IRT and the achievement scores specified according to the CTT.*

CAT Ability scores (IRT)	AVE
	.799**
Paper pencil test (CTT)	.000
	128

**. $p<0.01$

Table 3 shows that there is a high, positive, significant relationship between students' levels of knowledge that are estimated in CAT systems and the score of an achievement test they received from a PPT (r=,799, p<.01). The fact that the test was developed either according to CAT or the CTT did not lead to a meaningful change in student achievement scores.

3.3. Findings for the Third Sub-Purpose

As previously specified, 40 questions were asked to each of the students in a fixed length based on the CTT. In the CAT administered based on the IRT, a maximum of 40 questions and a standard error of 0,30 was taken as basis as a rule of conclusion. It is seen from Table 1 that the students answered at least 19 and at most 30 questions. Table 4 presents the average number of questions students had to answer in the CAT and CTT, the standard deviation of the test, and the average reliability coefficients.

Table 4. *The average number of questions asked in the administering of the CAT and the fixed number of questions asked based on the CTT*

	Item mean	Std. deviation	Reliability coefficients
CAT (IRT)	20.02	0.295	0.913
Paper-pencil test (CTT)	40	0.427	0.720

When Table 4 is examined, it is seen that the average number of questions students had to answer in the CAT was 20.02. Based on this, students answered approximately half of the number of questions they had to answer based on the CTT, and moreover, a knowledge level estimation was reached "at "high reliability".

4. RESULTS and DISCUSSION

In this study, a CAT system that provided for multiple-choice achievement tests to be given based on the IRT was designed. After that, the "office programs" course achievement test was conducted in the designed CAT system. According to the findings acquired in the study, the knowledge levels of students were estimated with high reliability by the achievement test performed in the CAT system. This finding shows that the feature of "measurement with high sensitivity and therefore high reliability with the adaptive test" (Kalender, 2011; Weiss, 1982; Wise and Kingsbury, 2000) specified in the literature is provided and it is similar to the literature (Cisar et al., 2016; Kuo et al., 2015; Özbaşı and Demirtaşlı, 2015; Petscher et al., 2017; Zhang and VanLehn, 2017).

In addition, a high level, positive and significant relationship was found between the students' knowledge levels estimated in the CAT system and the scores they got from the PPT. This finding shows that the scores and ability measurements calculated based on the two separate theories can rank individuals similarly and that there is no statistically significant difference in the results if one is used instead of the other. This result is very similar to the results of the research (Chao et al., 2015; Çelen and Aybek, 2013) examining the relationship between the 2 tests. Zhang and VanLehn (2017) found that CAT contributed

positively to students' learning outcomes. Liu and Zhao (2017) designed an English vocabulary learning system based on IRT and at the end of his experimental study, they found that the CAT was more successful by working with fewer words.

CAT systems provide high-precision measurements by asking questions appropriate to their level instead of questions that are well above or below their own knowledge level (Weiss, 1982; Wise & Kingsbury, 2000). When the parameters of the questions answered by the students and the estimated knowledge levels of the students are examined, this situation is clearly seen. As a matter of fact, it was seen that the predicted knowledge levels of the students and the difficulty levels of the questions were quite close to each other.

In the designed CAT system, the average number of questions that students had to answer was 20. The number of questions for the Paper Pencil test is 40. Similarly, Chalmers (2016) found that the CAT system he designed made more precise measurements with fewer questions. Petscher et al. (2017) conducted an experimental study aimed at increasing the reading-comprehension capabilities of students. As a result, he found that the students who took the CAT were given a smaller passage and had better reading comprehension ability compared to the students in the control group. Kaptan (1993) reduced the adaptive form of the paper-pencil form of the math test, which consisted of 50 questions, to 14 questions. As a result, it was found that the adaptive test system reduced the number of items used in estimating the level of knowledge by 70% compared to the paper-pencil test. Similarly, Gardner et al. (2004) similarly used 21 questions in the PPT form of the Beck depression scale, as opposed to an average of 6 questions in the adaptable version of the scale in a study they conducted. As a result, it can be said that the tests based on IRT measure all knowledge levels from low to high with high precision and significantly reduce the number of questions.

4.1. Suggestions

The CAT system can provide more reliable and effective results in courses that require problem-solving skills such as Programming Languages. Because students have the most difficulty in subjects that require problem solving skills. This may be due to the inability to develop an application suitable for each student's abilities. CAT tests can be applied frequently in the teaching process of subjects such as Science, Mathematics, Programming and Language learning. Feedback can be made for each student on necessary topics based on the results obtained from the CAT. It is obvious from this study and the literature that CAT systems are very advantageous. However, because the designing of CAT systems requires specialized knowledge, the number of studies conducted is quite low. This is why a graphic interface CAT platform can be designed that individuals who are not specialized in the topics of computer programming, measurement, and evaluation can easily use.

5. REFERENCES

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

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71*(5), 1-39. http://doi.org/10.18637/jss.v071.i05.

Chao, R.-C., Kuo, B.-C., & Tsai, Y.-H. (2015). Development of Chinese Computerized Adaptive Test System Based on Higher-Order Item Response Theory. *International Journal of Innovative Computing, Information and Control*, *11*(1), 57–76.

Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education*, *51*(2), 624–645. http://doi.org/10.1016/j.compedu.2007.06.011.

Cisar, S. M., Cisar, P., & Pinter, R. (2016). Evaluation of knowledge in Object Oriented Programming

course with computer adaptive tests c Cisar. *Computers & Education*, *93*, 142–160. http://doi.org/10.1016/j.compedu.2015.10.016.

Crocker, L., & Algina, J. (1986). *Introduction Classical and Modern Test Theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Çelen, Ü. & Aybek, E. C. (2013). Examination of the correlation between student achievements and the items that construct the test according to the Classical Test Theory and Item Response Theory. *Journal of Measurement and Evaluation in Education and Psychology*, *4*(2), 64–75.

Embretson, E. & Reise, S. P. (2000). *Item response theory for psychologist principles and application.* London: Lawrence Erlbaum Assc.

Gardner, W., Shear, K., Kelleher, K., Pajer, K., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, *4*(1), 13-23.

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 and Education*, *52*(1), 53–67. http://doi.org/10.1016/j.compedu.2008.06.007

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. Unpublished Doctoral Thesis. Ankara: ODTU.

Kaptan, F. (1993). C*omparison of the computerized, individual adaptive test method and paper-pencil test with respect to estimate the ability.* [Unpublished Doctoral dissertation, University of Hacettepe]. National thesis center of the Republic of Turkey.

Kuo, B. C., Daud, M., & Yang, C.-W. (2015). Multidimensional Computerized Adaptive Testing for Indonesia Junior High School Biology. *EURASIA Journal of Mathematics, Science & Technology Education*, *11*(5), 1105–1118. http://doi.org/10.12973/eurasia.2015.1384a.

Liu, Y., & Zhao, X. (2017). Design Flow of English Learning System Based on Item Response Theory. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(12), 91–102.

Lord, F.M., & Novick, M.R. (2008). *Statistical theories of mental test scores*. IAP.

Özbaşı, D., & Demirtaşlı, N. (2015). Development of Computer Literacy Test as Computerized Adaptive Testing. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 218–237.

Özyurt, H. (2013). *The Development and Evaluation of a Web Based Adaptive Testing System: The Case of Probability Unit*. Unpublished Doctoral Thesis. Trabzon: KTU.

Pelanek, R. (2016). Applications of the Elo rating system in adaptive educational systems nek. *Computers & Education*, *98*, 169–179. http://doi.org/10.1016/j.compedu.2016.03.017.

Petscher, Y., Foorman, B. R., & Truckenmiller, A. J. (2017). The Impact of Item Dependency on the Efficiency of Testing and Reliability of Student Scores from a Computer Adaptive Assessment of Reading Comprehension. *Journal of Research on Educational Effectiveness*, *10*(2), 408–423. http://doi.org/10.1080/19345747.2016.1178361.

Tseng, W.T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers and Education*, *97*, 69–85. http://doi.org/10.1016/j.compedu.2016.02.018.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer.* Routledge.

Weiss, D. (1982). “Improving Measurement Quality and Efficiency with Adaptive Testing.” *Applied Psychological Measurement*, *6*(4), 479–492. http://doi:10.1177/014662168200600408.

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

Wu, H. M., Kuo, B. C., & Wang, S. C. (2018). Computerized Dynamic Adaptive Tests with Immediately Individualized Feedback for Primary School Mathematics Learning. *International Forum of Educational Technology & Society*, *20*(1), 61–72.

Zhang, L., & VanLehn, K. (2017). Adaptively selecting biology questions generated from a semantic network. *Interactive Learning Environments*, *25*(7), 828–846. http://doi.org/10.1080/10494820.2016.1190939.