Uyarlamalı Bilgisayar Testlerinde Bulanık Mantık Kullanılması

Araştırma Makalesi/Research Article

Atila BOSTAN

Yazılım Mühendisliği Bölümü, Ankara Bilim Üniversitesi, Ankara, Turkiye
atila.bostan@ankarabilim.edu.tr

(Geliş/Received:03.01.2020; Kabul/Accepted:25.06.2020)
DOI: 10.17671/gazibtd.669791

Özet— Bu çalışmada, uyarlamalı test soru seçim algoritmalarında uzman bigisinden faydalanmanın katkısı incelenmiş ve bu yaklaşımın bilgisayar destekli uayarlamalı testlerde kullanılması sınanmıştır. İlave olarak çok boyutlu soruların ölçme başarısı üzerine etkileri de incelenmiştir. Bilgisayar destekli uyarlamalı testlerde uzman bilgisinden faydalanmak üzere bir algoritma önerilmiş ve deney ortamında test edilmiştir. Bulanık mantık hesaplama yöntemini kullanan, öğrenci başarı ölçüm algritması ile yedinci sınıf fen bilgisi dersinde bir durum çalışması gerçekleştirilmiştir. Sonuç olarak, bulanık mantık hesaplama yönetimini esas alan uyarlamalı bilgisayar testi ile yapılan ölçümlerin, öğrenci başarıları arasındaki farklılıkları daha ayırıcı şekilde vurguladığı istatistiksel yöntemler ile gösterilmiştir. Ayrıca, yedi farklı boyutu içeren uyarlamalı test uygulamasında öğrencilere en az 22, en çok 31 soru yöneltmek sureti ile yeterli doğrulukta bir değerlendirme yapılabildiği gözlenmiştir. Kendi kendine öğrenme ve uzaktan eğitim ortamlarında etkin olarak kullanılacağı değerlendirilen uyarlamalı testlerde, bulanık mantık hesaplaması kullanmanın uygun bir çözüm olabileceği tespit edilmiştir.

Anahtar Kelimeler— uyarlamalı bilgisayar testleri, bulanık mantık, çok-boyutlu sorular.

Using Fuzzy Logic in Computerized Adaptive Tests

Abstract— In this research, the contribution of expert knowledge in question selection algorithms of adaptive tests is studied and the employment of this approach in computerized adaptive tests is examined. In addition, the effect of multi-dimensional questions on measurement is also addressed. An algorithm utilizing the expert knowledge in computerized adaptive tests is proposed and tested in the experiments. A case study was conducted on seventh-grade science course with the utilization of fuzzy-logic based adaptive test. Consequently, by the help of the statistical studies, it is shown that computerized adaptive testing expresses the differences in student achievement levels in more visible way. Furthermore, 31 questions in maximum and 22 questions in minimum were observed to be effective for accurate assessment or student achievement levels on seven dimensions in adaptive testing. Using fuzzy logic in adaptive tests which were assessed to be effective in self-learning and distance education environments is found to be plausible.

Keywords— computerized adaptive tests, fuzzy logic, multi-dimensional questions.

1. INTRODUCTION

Teachers are expected to be mastered in human communication besides being a specialist in their respective teaching domain. Intrinsic to this expectation, teachers should be talented enough to communicate the context in teaching-domain and moreover, they should be capable to assess student achievement by utilizing a proper measuring technique [1]. In regular classroom environment, it may be straightforward for a teacher to evaluate student performance, if sufficient direct communication with student is provided. In other words, teachers could figure out a judgment value on student achievement level with the help of oral and face to face communication during classroom interactions [2][3]. When forming his/her judgment value, teacher intuitively evaluates the student responses to his/her insertions and questions. If needed, more explanatory (more specific) insertions or questions may follow to provide grounds for shaping a more precise judgment value. Face to face interactive achievement-assessments between teacher and students may be the most effective one, if the subjectivity which stems from personal factors such as; lack of expertise, insufficient communication (due to the number of the students in the class /disabilities), indefinite criteria and norms are excluded [2].

To refrain from the subjectivity influence on assessment, standardized tests are widely used in student evaluation processes. Carroll [4] pointed out that in standardized tests scoring method, examinees are supposed to provide a response for each of the presented items and the assigned weight for the item is used to measure the examinees' response. Although the utilization of recent technological means provides new and more sophisticated tools, Reeves [5] attributed the educational measurement as "*...still a relatively weak component of e-learning programs.*"

In Psychological Measurement Theory, first introduced by Baker and Harold [6] , they stated that "*All measurement is imprecise*" and they mentioned "*... since we are dealing with the mind, we will still remain in the land of inference and inevitably be left to piece together what has actually been experienced by the learner*". As can be deduced from these statements, it is impossible to provide a sharp and objective quantitative value on student achievements. Moreover, it is always impractical to use these strict values (anyhow attained from classical tests) without an accurate review and consideration.

> *"We must remember we are dealing with people not plastics. People are dynamic; they all change second to second. The meanings they ascribe to events become successively refined and restructured with experience. They are blurry targets for precise metrics."* [6].

It is widely accepted that the customization of educational progresses to meet the individual needs and preferences, constitutes one of the basic principles to improve the effectiveness of education. In a customized educational life cycle, it is inevitable that the educational measurement be personalized. The intensive usage of information processing technology in education in recent decades provided the foundation to customize educational measurement [7][8][9]. Although, until recently it was deemed inapplicable (especially for the number of students in regular classrooms), the adaptation of educational measurement to the individual attributes now found potential to exist with the help of the computing power, presentation alternatives and the interactive operations on wide range of databases which are provided by information processing technology. It is no doubt that the customized educational measurement would form the suitable basis for effective measurement on achieved individual learning [10].

Although, automatic adaptation of the teaching method to the students' learning style and requirements seems to be hard to reach, customizing student achievement measurement system to the student responses is more promising. In order to improve the applications on this hopeful path, in this research, the practicability of expert knowledge represented by fuzzy logic membership functions in computerized adaptive tests is examined. It is hoped that this kind of an algorithm would provide grounds to reduce the number of questions to be used in adaptive tests. In this research exploitation of multidimensional questions in computerized adaptive tests is also studied.

2. BACKGROUND

Adaptive measurement has gained much interest in recent years. Several large-scale testing programs, such as the Graduate Record Examination (GRE) and Test of English as a Foreign Language (TOEFL) have already switched from conventional paper and pencil version to computerized adaptive testing for the sake of efficiency and effectiveness [11]. By providing the opportunity to select the questions interactively for each student, the adaptive measurement approach is found effective in measuring the competency of student with respect to criteria, rather than comparing his/her achievements with that of classmates [1]. Adaptive measurement has received substantial scientific interest in recent years and the logic has been scientifically examined in selecting test questions [12][13], as well as competition systems [14]. All these studied reported significant advantages over classical measurements.

"*Paper-and-pencil tests are typically 'fixed-item' tests in which the examinees answer the same questions within a given test booklet. Since everyone takes every item, all examinees are administered to items that are either very easy or very difficult for them. These easy and hard items are like adding constants to someone's score…Examinees can be given the items that maximize the information (within constraints) about their ability levels from the item responses. Thus, examinees will receive few items that are very easy or very hard for them. This tailored item selection can result in reduced standard error and greater precision with only a handful of properly selected items*." [15].

Furthermore, by not been exposed to the questions which are too difficult for the student, the confidence and morale would be protected against corruption [16]. In this approach to educational measurement, the achievement level in learning processes is determined more effectively than classical standard tests, since the questions are chosen for each student to help to determine his/her achievement level particularly. Reducing the mean examination (measurement) time for a student and offering flexible measurement applications are among the benefits of this approach [17][18]. Whereas, the presence of the information processing infrastructure, the construction and maintenance of question database form the basic obstacles for the implementation of computerized adaptive tests [19] [20]. On the other hand, self-adequacy of adaptive testing approach in planning, repetition and interpretation of measurements suggest that it could be used effectively in personal education sets, such as distant education applications which lack real time student-teacher face to face interaction [21].

The question selection algorithm in computerized adaptive tests depends heavily on Item Response Theory (IRT)[22]. In IRT, an ability scale metric is to be developed by applying the test to a group of examinees whose ability level is identified and known. This, needs the items in a test should be exposed to several test runs before actual usage in a computerized adaptive test [11].

The importance of teacher's in-class assessments and measurements conducted by oral examinations can not be refuted in classical classroom teaching applications. Especially oral examinations are practical examples of customized adaptive measurement [3]. When oral examinations are investigated with the view of student-teacher interaction, it could obviously be determined that the teacher's level of experience in oral examination plays a great deal of importance. In this type of measurements, teacher should be capable of deriving an evaluation on student's achievement level, besides bearing sufficient administrative skills on interactive adaptive measurement [1]. As in most of the expert systems, the algorithm of evaluation which is used by teaches in oral examinations cannot be formulized in classical mathematical terms. When administering the oral examination, teacher benefits from the emotional interactions, body language, experiences gained before the examination and some other factors which are hard to be formulized. In forming mathematical formula for this kind of an interaction, not only the complexity of model is an obstacle but also the impossibility in quantifying most of the effective dimensions. Particularly, in similar problems where expert knowledge and skill have great influence on the outcome of the process, fuzzy logic calculation has presented considerable success [23]. By defining fuzzy values for the attributes that are not possible to quantify or are not needed to be stated in quantities, the expert functioning can be more effectively imitated. The experience of the teacher has enormous effect on determination of student achievement level. Handling of these above mentioned factors (which are advantaged by teachers) by automated systems will provide substantial contribution to the educational measurement.

In recent years, some researchers focused on application of fuzzy set theory in to education. Ma and Zhou [24] studied applying fuzzy set approach to student-centered learning assessment. Ibrahim [25] proposed a fuzzy system for evaluating students' learning achievement. Shih-Ming [26] presented an automatic normalization of lenient-grades by using fuzzy set membership functions. In recent years, several scientific studies have focused on utilization of fuzzy computation in improving the effectivity of educational measurement and environments [27][28][29]. There are several different calculation strategies in fuzzy logic with respect to the application environment. Determination of success level for different fuzzy calculation strategies in educational measurement will put a light on the future experiments in this realm.

It is ideal that each question should measure single dimension. But, the correlation among the student achievements (especially in some education domains such as mathematics and science) is inevitable in most of the cases [30]. The practicability of measuring single dimension heavily depends on the assumption that the prerequisites are been successfully achieved. However, this assumption causes the misinterpretation of wrong answers, if the reason which lies behind the wrong answer is other than the measurement dimension. Another drawback with this assumption is that, although each question in a regular test application is related with several dimensions, only one of them is blamed for the result. In his instructional documentation, Rudner [31] mentioned the rationale behind the Item-Analysis methodologies as the probability that the questions in a test may be more related with other dimensions than the targeted dimension. If the underlying actual cause(s) for a wrong answer could be identified among the group of correlated ones, a new and prosperous contribution to the educational measurement would be provided.

3. METHODOLOGY

To elicit the effectiveness of fuzzy logic in selecting adaptive test questions, a pre-test and post-test research model is employed on three experimental student groups at seventh-grade of secondary school.

The secondary school in which the experiment is conducted is subjectively chosen for the reason that the school has a suitable computer laboratory and the students have developed a satisfactory level of computer usage skill. Since the experimental model utilizes pre-test and post-test model and the applied tests were individually customized, it is evaluated that the subjectivity in selection of the school would not affect the experiment results.

The experimental implementation took place in Cağribey Secondary School, located in Keçioren district of city of Ankara-Turkey. All three sections (A,B,C) of seventh grade were included in the experiment. Also in the experiment, students were divided into three application groups; however, to assure the randomization of student distribution among the experiment groups, the allocation of students into the groups was done on the alphabetic order of student names.

Each application group was composed of 38 students, a total of 114 students were incorporated in the experiment. The first group was the control group and took both pre-test and post-test in the conventional form. The second group (experimental group 1) took the customized adaptive test beginning with moderate difficulty level questions in the post-test. On the other hand, the third group (experimental group 2) took customized adaptive test beginning with student selected difficulty level questions in the post-test.

In the examination of the test results ANNOVA, paired samples t-test, Shaffe test, correlation level, average and percent values are presented, compared and discussed where appropriate

4. EXPERIMENTS AND DATA COLLECTION

In determination of coefficients for questions and fuzzy membership values for answer options, expert evaluations were used. Expert assessments were acquired with the help of face to face inquiries, group discussions and questionnaire applications.

The competencies which were chosen to be measured in this experiment are listed in Table 1 below. The numbers are assigned to the competencies in order to facilitate representation in Table 1. The questions which were used in the experiments are directly linked with the competencies stated in Table 1.

Table 1. Competencies

Unit	Competency	
	Nun	Explanation
Force & Motion	1	To understand the relation between force and motion.
	2	To comprehend the mathematical equation on relation among velocity, time and distance. Practice on problems using these relations.
	3	To compute vectorial quantities (force and motion).
Simple Machines	4	To understand the principles of lever mechanisms. Solving problems using the relation among force, lever arms (force and load) and load.
	5	To understand the principles of belt pulley mechanism. Solving belt pulley mechanism problems.
	6	To understand the principles of fixed and movable pulley systems. Solving pulley problems.
	7	To understand the similarities between lever and winding wheel mechanisms. Solving problems in winding wheel mechanisms.

In the pre-test all three groups answered 50 questions. Pre-test was applied simultaneously to all three groups and the test included 50 questions which were composed of 7 questions for each one of the first 6 competencies and 8 questions for the last one.

In the post-test, the control group received another 50 question test while other two group were subject to the adaptive testing. Post-test questions for control group were chosen among the questions which were used in adaptive testing of the other two experimental groups. In both pre-test and post-test time was not controlled for the test duration. Test was concluded when the students finished the application.

4.1. Assigning Fuzzy Membership Values and Fuzzy Inference System Selection.

The interrelations between student achievement level and fuzzy sets were mapped using fuzzy membership values. With the help of the inquiries, discussions and questionnaire applications, identified fuzzy set definitions and graphs are shown in Table 2 and Figure 1.

Although in fuzzy logic implementations six membership types were common, namely triangular, trapezoidal, Gaussian, generalized Bell, Π-shaped and S-shaped, with their specific advantages and disadvantages reported [32][33], the selection of the membership type was not an option in this study since membership degrees of the achievement grades to fuzzy sets were constructed by the views of the experts. Eventually, the data collected in our study was best expressed by trapezoidal membership type. This finding may be specific to the educational field or science education in particular.

Table 2. Boundaries for fuzzy sets

Fuzzy Set	LB (0)	LB (1)	HB (1)	HB (0)
Very-Low	-	0	9	14
Low	8	18	31	43
Moderate	24	36	60	73
High	63	69	81	92
Very-High	79	83	100	-

LB (x) : Lowest boundary for value “x”
HB (x) : Highest boundary for value “x”

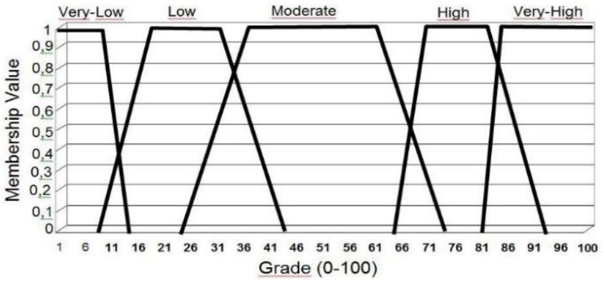

Figure 1. Fuzzy set boundaries

For the difficulty metric of questions five different fuzzy sets (very easy, easy, moderate, somewhat hard, hard) were used.

In fuzzy inference computation, Mamdani type of a fuzzy inference mechanism is used in the study. Since the quantification of the student achievement-level given the answers to the questions is hard to be computed linearly. In addition, expert assessments were observed to be fluctuating (as a crisp value) on the quantity of student-achievement level when a question is answered correctly, but the variation between the assessments was small enough to express the achievement as a fuzzy set. With these observations, Mamdani type of a fuzzy computation was evaluated to be more suitable to the problem, because the other common alternative to fuzzy inference mechanism, Sugeno type of fuzzy computation, is good on scalar, whereas Mamdani is on fuzzy output calculation [34][35].

4.2. Algorithm for Question Selection

The questions which have multiple relations with the competencies that are selected for measurement have precedence over the others. In application some threshold values were used; to stop testing, for sufficiency and insufficiency of displayed achievement and for minimum pace. Minimum number of questions to judge on an achievement level was chosen to be 2, and the maximum number of questions which can be asked on a given competency was applied as 5. The algorithm which utilizes these parameters is stated below.

1. Identify the competency to be measured and assign them to a set (namely dimensions-set).

2. Assign first competency as measurement-focus. If the dimensions-set is empty, then jump to step 9.
3. Set difficulty-level indicator as “moderate”.
4. Identify the questions which have relation with the competency in focus with the magnitude pointed by difficulty-level indicator.
5. Discard the asked (previously used/directed questions in the same test) questions.
6. Chose the one which has highest relation with the competency in focus. (the question which has high number of relations with the competencies in dimensions-set is preferred)
7. Direct the question to the student and store the response.
8. Add fuzzy membership value of the student response to the proper achievement pointer to calculate the measured success level for each competency.

   8.1.1. If the maximum number of questions for the competency in focus is reached, or upper-threshold value for the competency success level is reached for the competency in focus then remove it from the dimensions-set and return to step 2.

   8.1.2. If the upper-threshold value for the difficulty-level membership is reached then difficulty-level pointer is incremented to one step higher, else if lower-threshold value for difficulty-level membership is reached then difficulty-level pointer is decremented to one step lower.

   8.1.3. The displayed improvement in measured sufficiency is compared with the minimum pace.

   8.1.3.1. If equal or displayed improvement is greater than minimum pace, then jump to step 4.

   8.1.3.2. If displayed improvement is less than minimum pace, then remove the dimension in focus from the dimensions-set and return to step 2
9. Summed up sufficiency membership values for each of the competencies in dimensions-set are converted to competency sufficiency points by using defuzzification algorithms.

5. RESULTS

In this section, results observed in the experiments are reported with significant interpretations on them. Initially, pre-test results for all three experimental groups are presented and subsequently, in order to verify the equivalency of the post-tests, post-test result for the control group is elicited. Later, the effect of fuzzy logic calculation in adaptive testing is investigated. Additionally, for defuzzification of the test results two different approaches, namely center-of-gravity and center-of-maximum-value, are compared for their efficiency. Total number of questions asked discreetly in each experiments also compared and discussed as well. Finally, results and interpretation on the effect of multidimensional question in adaptive testing are highlighted.

5.1. Pre-test and Equivalency of the Post-test

Three groups took the pre-test to examine the random distribution of students to the groups and to test the differences in gained achievement levels among the groups. The results of ANOVA test are shown in Table 3.

As can be seen in Table 3 there is no difference (with confidence 0,05) in gained achievement levels among the three groups on the basis of seven measurement targets. The control group took a second fixed 50-question test to examine equity of pre-test and post-test question batteries and also to test the effects of a second test on the same topics following the pre-test, and the test application time.

Table 3. ANNOVA test results of pre-test.

		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Pre-test Cmptncy 1	Bw Grps	1,947	2	,974	1,195	,307
	Win Grps	90,474	111	,815		
	Total	92,421	113			
Pre-test Cmptncy 2	Bw Grps	,491	2	,246	,406	,667
	Win Grps	67,132	111	,605		
	Total	67,623	113			
Pre-test Cmptncy 3	Bw Grps	1,491	2	,746	1,232	,243
	Win Grps	163,000	111	1,468		
	Total	164,491	113			
Pre-test Cmptncy 4	Bw Grps	5,333	2	2,667	1,709	,186
	Win Grps	173,158	111	1,560		
	Total	178,491	113			
Pre-test Cmptncy 5	Bw Grps	1,070	2	,535	,286	,752
	Win Grps	207,421	111	1,869		
	Total	208,491	113			
Pre-test Cmptncy 6	Bw Grps	,333	2	,167	,114	,892
	Win Grps	161,632	111	1,456		
	Total	161,965	113			
	Bw Grps	,211	2	,105	,068	,935
	Win Grps	172,421	111	1,553		

Pre-test	Total	172,632	113			
Pre-test Grand Total	Bw Grps	17,070	2	8,535	,798	,453
	Win Grps	1187,18	111	10,69		
	Total	1204,25	113			

Bw Grps: Between Groups **Win Grps:** Within Groups

The data are reviewed with Paired Samples T-Test. Test result is shown in Table 4. It is evident in table, no significant difference with confidence 0,05 is observed among the displayed achievement levels between pre-test and post-test application in all seven competencies and total views.

Table 4. Paired sample test results.

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Dvtn	Std. Error Mean	95% Cnfdnce Interval of the Difference				
				Lower	Upper			
Pr1 – Ps1	-,053	1,451	,235	-,530	,424	-,224	37	,824
Pr2 – Ps2	,105	,798	,130	-,157	,368	,813	37	,422
Pr3 – Ps3	,079	1,239	,201	-,328	,486	,393	37	,697
Pr4 – Ps4	,026	1,026	,166	-,311	,364	,158	37	,875
Pr5 – Ps5	-,105	,953	,155	-,418	,208	-,681	37	,500
Pr6 – Ps6	,053	1,064	,173	-,297	,402	,305	37	,762
Pr7 – Ps7	,132	1,044	,169	-,212	,475	,777	37	,442
PrT - PsT	,290	2,588	,420	-,561	1,140	,690	37	,495

Pr/PsX: Pre/Post-Test Competency X
Pr/PsT: Pre/Post test grand total

In order to test that pre and post-tests are measuring the same dimensions, the correlation indexes are given in Table 5. A strong relation between the two tests is evident in Table 5 with regards to correlation coefficients having values between 0,57-0,74. These strong relations can be considered as a significant indicator for functional equivalency of these two tests.

Table 5. Correlation test results between pre and post tests

.

	N	Correlation	Sig.
Pr1 – Ps1	38	,659	,007
Pr2 – Ps2	38	,568	,000
Pr3 – Ps3	38	,690	,006
Pr4 – Ps4	38	,679	,000
Pr5 – Ps5	38	,742	,001
Pr6 – Ps6	38	,686	,000
Pr7 – Ps7	38	,664	,000
PrT – PsT	38	,730	,000

Pr/PsX: Pre/Post-Test Competency X
Pr/PsT: Pre/Post test grand total

Presented analysis results specify that a comparison between pre and post-test results would not lead to an erroneous conclusion.

5.2. The Effect of Fuzzy Logic Calculation in the Measurement of Student Achievements.

The count of wrong and correct answers given by the two experimental groups for each of the questions in seven competencies are presented in Table 6.

Table 6. Correct and wrong answer counts.

Competency	Question #	Experimental Group 1						Experimental Group 2					
		C	**W**	**Count**	**Suc. %**	**Tot.Q.**	**Ave. Suc. %**	**C**	**W**	**Count**	**Suc. %**	**Tot.Q.**	**Ave. Suc. %**
1	**1**	35	3	38	92,11	124	87,90	33	5	38	86,84	133	84,21
	2	33	5	38	86,84			33	5	38	86,84		
	3	24	5	29	82,76			22	8	30	73,33		
	4	11	1	12	91,67			16	1	17	94,12		
	5	6	1	7	85,71			8	2	10	80,00		
2	**1**	33	5	38	86,84	131	86,26	32	6	38	84,21	130	84,62
	2	31	7	38	81,58			30	8	38	78,95		
	3	25	6	31	80,65			23	6	29	79,31		
	4	17	0	17	100,00			18	0	18	100,00		
	5	7	0	7	100,00			7	0	7	100,00		
3	**1**	28	10	38	73,68	148	79,05	29	9	38	76,32	148	81,08
	2	30	8	38	78,95			28	10	38	73,68		
	3	27	9	36	75,00			30	6	36	83,33		
	4	22	0	22	100,00			21	0	21	100,00		
	5	10	4	14	71,43			12	3	15	80,00		
4	**1**	24	14	38	63,16	161	69,57	22	16	38	57,89	159	68,55
	2	21	17	38	55,26			23	15	38	60,53		
	3	27	11	38	71,05			27	11	38	71,05		
	4	25	6	31	80,65			26	7	33	78,79		
	5	15	1	16	93,75			11	1	12	91,67		
5	**1**	24	14	38	63,16	165	43,64	29	9	38	76,32	159	42,77
	2	17	21	38	44,74			14	24	38	36,84		
	3	14	24	38	36,84			11	27	38	28,95		
	4	12	23	35	34,29			10	25	35	28,57		
	5	5	11	16	31,25			4	6	10	40,00		
6	**1**	11	27	38	28,95	155	27,74	8	30	38	21,05	153	32,03
	2	12	26	38	31,58			10	28	38	26,32		
	3	11	27	38	28,95			15	23	38	39,47		
	4	6	20	26	23,08			10	14	24	41,67		
	5	3	12	15	20,00			6	9	15	40,00		
7	**1**	8	30	38	21,05	147	15,65	11	27	38	28,95	147	17,01
	2	4	34	38	10,53			4	34	38	10,53		
	3	10	28	38	26,32			9	29	38	23,68		
	4	1	18	19	5,26			1	19	20	5,00		
	5	0	14	14	0,00			0	13	13	0,00		

C : Correct answer count **W :** Wrong answer count
Count : Question count **Suc. % :** Success percentage
Tot.Q.: Total question count for competency
Ave.Suc.% : Average success percentage for competency

The data obtained from the analysis of variation among the post-test results of three groups (control and two experimental groups) is shown in Table 7. When Table 7 is studied, it is determined that there is significant difference among the success levels of three groups in the basis of seven dimension groups. But in grand total, the difference among the groups is not significant.

Table 7. ANNOVA test results of post-test.

		Sum of Squares	df	Mean Square	F	Sig.
Post-test Cmptncy 1	Bw Grps	12948,620	2	6474,310	7,713	,001
	Win Grps	93174,537	111	839,410		
	Total	106123,157	113			
Post-test Cmptncy 2	Bw Grps	25162,237	2	12581,119	16,431	,000
	Win Grps	84991,858	111	765,692		
	Total	110154,095	113			
Post-test Cmptncy 3	Bw Grps	7878,848	2	3939,424	4,005	,015
	Win Grps	93958,836	111	846,476		
	Total	101837,684	113			
Post-test Cmptncy 4	Bw Grps	6581,689	2	3290,844	3,832	,025
	Win Grps	95328,545	111	858,816		
	Total	101910,234	113			
Post-test Cmptncy 5	Bw Grps	2081,790	2	1040,895	3,784	,042
	Win Grps	106551,911	111	959,927		
	Total	108633,701	113			
Post-test Cmptncy 6	Bw Grps	20433,848	2	10216,924	12,895	,000
	Win Grps	87943,784	111	792,286		
	Total	108377,632	113			
Post-test Cmptncy 7	Bw Grps	22699,275	2	11349,638	12,986	,000
	Win Grps	97014,571	111	874,005		
	Total	119713,846	113			
Post-test Grand Total	Bw Grps	75,236	2	37,618	,867	,423
	Win Grps	4817,032	111	43,397		
	Total	4892,268	113			

To clarify reasoning for the differentiation among the groups a Scheffe test is conducted and the results of this test is presented in Table 8.

As it is seen in Table 8, the differences between the control group and both experimental groups are significant in each competency. Meaning that control-group results in the post-test are significantly different than that of both experiment groups in each competency. However, no significant differentiation between the two experimental groups is observed. In other words, the observed difference between post-test results of two experiment groups in each competency is not statistically significant. So that we cannot say post-test results of experiment groups are different in any competency.

On the other hand, in grand-total comparisons, it is seen that there is no significant difference in any group pairing. This is an interesting finding which can be explained with the neutralization of the observed differences in each competency when the test is assessed as a single measurement or competencies are not taken into account. This means the differences in the competency level diminish each other when the post-test result is assessed as a cumulative grade per student.

Table 8. Scheffe test results between experiment groups.

Dependent Variable	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Post-test Cpt 1	Cnt Gr.	Exp-1	23,939(*)	6,647	,002	7,447	40,430
		Exp-2	20,988(*)	6,647	,008	4,496	37,480
	Exp-1	Cnt Gr.	-23,939(*)	6,647	,002	-40,430	-7,447
		Exp-2	-2,951	6,647	,906	-19,442	13,541
	Exp-2	Cnt Gr.	-20,988(*)	6,647	,008	-37,480	-4,496
		Exp-1	2,951	6,647	,906	-13,541	19,442
Post-test Cpt 2	Cnt Gr.	Exp-1	34,932(*)	6,348	,000	19,181	50,683
		Exp-2	26,301(*)	6,348	,000	10,550	42,052
	Exp-1	Cnt Gr.	-34,932(*)	6,348	,000	-50,683	-19,181
		Exp-2	-8,631	6,348	,400	-24,382	7,120
	Exp-2	Cnt Gr.	-26,301(*)	6,348	,000	-42,052	-10,550
		Exp-1	8,631	6,348	,400	-7,120	24,382
Post-test Cpt 3	Cnt Gr.	Exp-1	9,185(*)	7,013	,017	8,215	26,586
		Exp-2	7,892(*)	7,013	,033	9,508	25,293
	Exp-1	Cnt Gr.	-9,185(*)	7,013	,017	-26,586	-8,215
		Exp-2	-1,293	7,013	,983	-18,693	16,108
	Exp-2	Cnt Gr.	-7,892(*)	7,013	,033	-25,293	-9,508
		Exp-1	1,293	7,013	,983	-16,108	18,693
Post-test Cpt 4	Cnt Gr.	Exp-1	6,911(*)	6,723	,019	9,770	23,593
		Exp-2	18,422(*)	6,723	,026	1,740	35,103
	Exp-1	Cnt Gr.	-6,911(*)	6,723	,019	-23,593	-9,770
		Exp-2	11,510	6,723	,235	-5,171	28,191
	Exp-2	Cnt Gr.	-18,422(*)	6,723	,026	-35,103	-1,740
		Exp-1	-11,510	6,723	,235	-28,191	5,171
Post-test Cpt 5	Cnt Gr.	Exp-1	-7,369(*)	7,108	,038	25,005	10,267
		Exp-2	2,754(*)	7,108	,029	-14,882	-20,390
	Exp-1	Cnt Gr.	7,369(*)	7,108	,038	-10,267	-25,005
		Exp-2	10,123	7,108	,366	-7,513	27,759
	Exp-2	Cnt Gr.	-2,754(*)	7,108	,029	20,390	14,882
		Exp-1	-10,123	7,108	,366	-27,759	7,513
Post-test Cpt 6	Cnt Gr.	Exp-1	-24,090(*)	6,458	,001	-40,112	-8,068
		Exp-2	7,226(*)	6,458	,037	-8,796	-23,248
	Exp-1	Cnt Gr.	24,090(*)	6,458	,001	8,068	40,112
		Exp-2	31,316	6,458	,566	15,294	47,338
	Exp-2	Cnt Gr.	-7,226(*)	6,458	,037	23,248	8,797
		Exp-1	-31,316	6,458	,566	-47,338	-15,294
Post-test Cpt 7	Cnt Gr.	Exp-1	-28,657(*)	6,782	,000	-45,485	-11,829
		Exp-2	17,408(*)	6,782	,039	14,420	19,236
	Exp-1	Cnt Gr.	28,657(*)	6,782	,000	11,829	45,485
		Exp-2	31,065	6,782	,610	14,237	47,893
	Exp-2	Cnt Gr.	-17,408(*)	6,782	,039	-19,236	-14,420
		Exp-1	-31,065	6,782	,610	-47,893	-14,237
Grand Total	Cnt Gr.	Exp-1	-,974	1,511	,813	-4,723	2,776
		Exp-2	1,016	1,511	,798	-2,734	4,766
	Exp-1	Cnt Gr.	,974	1,511	,813	-2,776	4,723
		Exp-2	1,990	1,511	,423	-1,760	5,740
	Exp-2	Cnt Gr.	-1,016	1,511	,798	-4,766	2,734
		Exp-1	-1,990	1,511	,423	-5,740	1,760

Cpt : Competency **Cnt Gr. :** Control Group
Exp-x: Experiment Group x

It is apparent that the control group results for two tests are very close to each other when examined in per competency and in group total. But in both experimental groups the difference between the two tests are significant and in the way to emphasize the success level. In other words, it could be stated that the applied algorithm in computerized adaptive testing expressed the differences in student achievement levels in a more visible way. The data lead to this interpretation is shown in Table 9.

Table 9. Span of the success levels in pre and post tests.

		Maximum		**Minimum**		**Diff.**
		Cpt. No.	**Suc. %**	**Cpt. No.**	**Suc. %**	
Pre-test	Control Gr.	1	78,57	7	42,11	36,47
	Expt.Gr.1	1	80,45	7	41,45	39,00
	Expt.Gr.2	2	80,83	7	40,79	40,04
Post-test	Control Gr.	2	77,07	7	40,46	36,61
	Expt.Gr.1	1	87,90	7	15,65	**72,26**
	Expt.Gr.2	2	84,62	7	17,01	**67,61**

Expt.Gr.x: Experimental group x **Suc.% :** Success percentage
Diff.: Difference **Cpt. :** Competency

In pre-test, the maximum difference between the highest and the lowest sufficiency levels in the second experimental group with 40,04 points. Whereas, the maximum difference between the highest and the lowest sufficiency levels in post-test is in the first experimental group with 72,26 value. These differences between maximum and minimum values (span) indicate that the proposed testing methodology is more capable to differentiate among student sufficiency levels.

5.3. Plausible Defuzzification Algorithm in Achievement Measurement

In defuzzification of student achievement levels both the Center of Gravity and Center of Maximum Value algorithms were applied and the results were compared with the pre test results on the basis of seven dimensions using Paired T-test comparison methodology. The comparison table for the Center of Maximum Value calculation results is given in Table 10.

Table 10. Paired sample test results of pre-test and center of mean defuzzification value.

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Dvtn	Std. Error Mean	95% Cnfdnce Interval of the Difference				
				Lower	Upper			
Pr1–CG1	30,107	33,643	5,458	19,049	41,165	5,517	37	,000
Pr2–CG2	32,331	36,501	5,921	20,333	44,329	5,460	37	,000
Pr3–CG3	34,403	37,500	6,100	17,956	36,762	2,722	37	,035
Pr4–CG4	5,191	42,834	6,949	-8,888	19,270	,747	37	,460
Pr5–CG5	,507	46,172	7,490	-14,670	15,684	,068	37	,946
Pr6–CG6	-21,948	37,495	6,083	-34,272	-9,623	-3,608	37	,001
Pr7–CG7	-31,600	32,981	5,350	-42,441	-20,759	-5,906	37	,000
PrT–CGT	,086	,829	,134	-,187	,358	,638	37	,527

Prx: Pre–test on Competency x **PrT:**Pre-test grand total
CMx: Center of Maximum Value on Competency x
CMT: Center of Maximum Value grand total

In Table 10, it is shown that the difference between the pre and post test results (when post test results were calculated with the Center of Maximum Value algorithm) is not significant in 0,05 confidence interval for the experimental groups.

The Correlation test results between the Center of Maximum Value calculation and the pre-test results are shown in Table 11.

Table 11. Correlation test results between pre-test and center of maximum defuzzification value.

	N	**Correlation**	**Sig.**
Pr1-CM1	38	,679	,003
Pr2-CM2	38	,588	,158
Pr3-CM3	38	,527	,010
Pr4-CM4	38	,674	,043
Pr5-CM5	38	,591	,007
Pr6-CM6	38	,416	,192
Pr7-CM7	38	,623	,026
PrT-CMT	38	,792	,000

Prx: Pre–test Competency x **PrT:** Pre-test grand total
CMx: Center of Maximum Value on Competency x
CMT: Center of Maximum Value grand total

The correlation coefficient between the two different computations were found at moderate level (0,41-0,68). Although, in only one of the dimensions (6th) the coefficient of correlation was weak, the coefficient calculated for the total pairs was 0.792, close enough to the strong relation categorization in correlation.

As for Center of Maximum Value, the Paired T test was conducted between the Pre-test and Center of Gravity algorithm results. The results for Paired T test is presented in Table 12.

As shown in Table 12, the difference between the results of pre-test and Center of Gravity calculation are significant in five dimension pairs (pairs 1,2,3,6, and 7).

Table 12. Paired sample test results of pre-test and center of gravity defuzzification value.

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Dvtn	Std. Error Mean	95% Cnfdnce Interval of the Difference				
				Lower	Upper			
Pr1–CM1	3,642	22,514	3,652	-3,758	11,042	,997	37	,325
Pr2–CM2	1,725	20,156	3,270	-4,900	8,350	,528	37	,601
Pr3–CM3	1,989	22,752	3,691	-5,489	9,467	,539	37	,593
Pr4–CM4	-2,810	22,432	3,639	-10,183	4,563	-,772	37	,445
Pr5–CM5	4,970	25,181	4,085	-3,307	13,246	1,217	37	,231
Pr6–CM6	2,921	29,284	4,751	-6,705	12,546	,615	37	,542
Pr7–CM7	1,095	25,109	4,074	-7,158	9,348	,269	37	,790
PrT–CMT	-,057	,753	,123	-,305	,191	-,467	37	,644

The Correlation test results between the Center of Gravity calculation and the pre-test results are shown in Table 13.

Table 13. Correlation test results between pre-test and center of gravity defuzzification value.

	N	Correlation	Sig.
Pr1-CG1	38	,068	,685
Pr2-CG2	38	-,316	,054
Pr3-CG3	38	,045	,790
Pr4-CG4	38	-,181	,276
Pr5-CG5	38	,318	,052
Pr6-CG6	38	,027	,874
Pr7-CG7	38	,144	,388
PrT-CGT	38	,728	,000

Prx: Pre–test Competency x **PrT:** Pre-test grand total
CMx: Center of Maximum Value on Competency x
CMT: Center of Maximum Value grand total

Even though the correlation coefficient for total pair (pair 8) indicates moderate relation strength, there are weak relations between the dimension pairs (pairs 1-7). Moreover, in pairs 2 and 4 the coefficient indicates a negative relationship.

With the findings in comparisons between pre-test and two different defuzzification algorithm results, it can be pointed out that the Center of Maximum Value algorithm produces more reasonable results.

5.4. Number of the Questions Asked

In the first experimental group a total of 1031 questions were used in the interactive test (some of the questions were asked more than one student). The distribution of these questions into the seven dimensions alone with the number of correct answers received is shown in Figure 2.

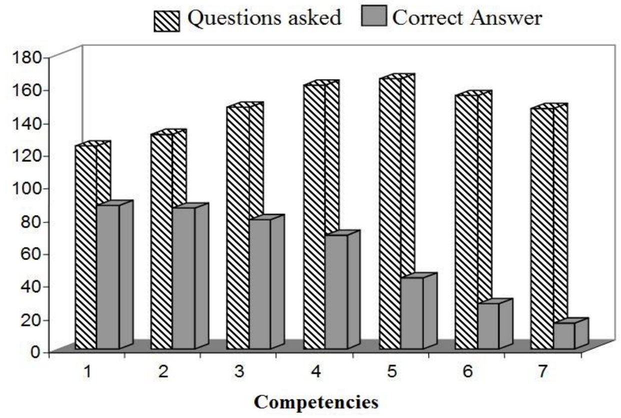

Figure 2. Number of questions asked and correct answers in experimental group 1

As can be seen in Table 6, the highest numbers for questions asked were observed in 4th and 5th dimensions with the values 161 and 165 respectively and the average number of correct answers received for these dimensions were calculated as 69.57 and 43.64 respectively. For the 1st dimension, in which the average number of correct answer was the highest (87.90), the total number of questions asked was observed as the lowest with the value 124. Also for the 7th dimension, in which the average number of correct answer was the lowest (15.65), the total number of questions asked was observed as 147.

In the second experimental group a total of 1029 questions were asked. The distribution of these questions into the seven dimensions alone with the correct answers received is shown in Figure 3. As it is observed in the first experimental group, the observed highest number for questions asked was in 4th and 5th dimensions with the value 159. The average numbers of correct answer in these groups were 68.55 and 42.77 respectively. For the 2nd dimension in which the highest number of correct answer was observed (84.62), the total number of questions asked was the lowest with value 130. In the 7th dimension in which the lowest number of correct answer was observed (17.01), the total number of questions asked was 147.

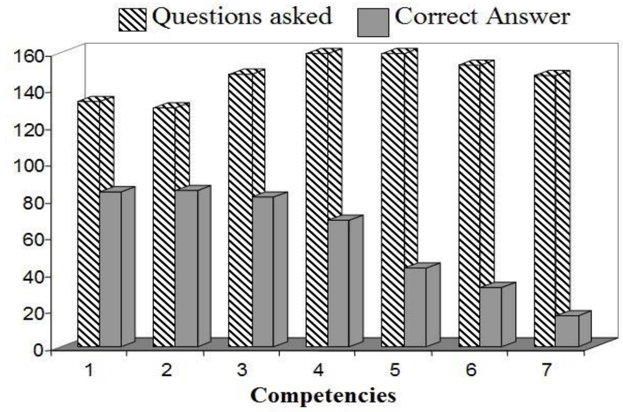

Figure 3. Number of questions asked and correct answers in experimental group 2

As for the algorithm used, at least 2 questions were asked in each dimension to decide on the student's achievement level. Therefore, in each of seven dimensions 1st and 2nd questions were asked by design to all students but succeeding questions (3^{rd},4^{th},5^{th}) were asked depending on the computed achievement level of the student in that dimension. Subsequently, it is observed that the exploitation number for the 3rd,4th and 5th questions increases in the dimensions in which the student's achievement level is found at moderate level (around %50 of success level), while exploitation number for these questions decreases as the student's achievement level closes to both the top and the lowest values of achievement measurement in a dimension. This an expected outcome for adaptive testing methodology; as the success level of an examinee is more predictable the number of questions used to determine the success level decreases, in other words for uncertain conditions you have to exploit more questions.

It is apparent in Figure 2, for the first experimental group the 3rd,4th and 5th questions in 4th and 5th dimensions were asked more than the other dimensions. By no surprise, the displayed success levels of these to dimensions are close to the median value.

The same effect is observed in the second experimental group. In the 4th and 5th dimensions in which the calculated success level was close to the middle value the 3rd,4th and 5th questions were asked more than the other dimensions. This is shown in Figure 3.

The maximum number of questions asked to an examinee is found as 31, while the minimum number is found as 22. The average number of questions asked in two of the experimental groups is 27.

5.5. Selection of the Initial Difficulty Level by Students

Students in the second experimental group had selected the initial difficulty level in the adaptive test while the first experimental group started with moderate level by default. In Table 8 the Scheffe test results are listed. It can be seen that the difference is not significant between the post test results of first and second experimental groups. Moreover, total numbers of the questions asked in first and second experimental groups are very close to each other with the values 1031 and 1029 respectively.

Each student in the second experimental group had selected the initial difficulty level in the adaptive test. But only 16 students out of 38 had chosen a different level other than middle. Nine of these 16 students' estimates were found to be inaccurate. Unfortunately, in the interviews conducted after the experiment, a great majority of the students declared that they had chosen the middle level for that they hadn't feel confident about their level of success in such a computerized adaptive test. Since this was the first time that they were subject to computerized adaptive testing. With the help of these interviews it is derived that the experiment needs to be performed again after the students develop consciousness in selecting the initial difficulty level for the questions in adaptive testing.

5.6. Multidimensional Questions

Average number of the questions asked in a dimension is shown in Table 14 for the three possible conditions. If there was no priori assessment value for the dimension in question, then the average number of the questions asked to reach judgment was observed as 3.87. If the priori assessment was positive or negative, then the number of the questions asked was observed as 3.26 and 4.34 respectively.

Table 14. Priori assessment and count of questions asked.

Priori Assessment Value	**Average count of Questions Asked in a Dimension**
Not Exists	3,87
Positive	3,26
Negative	4,34

As can be deduced from the figures the number of the questions asked is increasing when there exist a positive priori assessment and decreasing whet the priori assessment is negative.

When multi-dimensional questions are examined from the point of their contribution to the assessment process, it is found that the positive priori value decreases the number of the questions asked by an average value of 0.61 and negative value increases that number on average by 0.47. Consequently, it is deemed necessary that the study should be repeated with a different set of expert group and it should be extended to cross domain experiments to determine the contribution in cross domain dimensions.

5. CONCLUSIONS AND DISCUSSION

Integrating the fuzzy logic computation in computerized adaptive test applications has received considerable scientific attention. Because of their potential value in educational settings where the interaction between the teacher and student is limited, such as in distant education or customized tutorial applications. There are several scientific studies examining the fuzzy logic computation in automatic adaptive testing environments.

In their study Suarez-Cansino, and Hernandez-Gomez [36] studied a fuzzy inference mechanism for adaptive tests. They have simulated question selection mechanism on a database of questions that were labeled as high and low for their respective complexity level. In addition, they have classified the simulated students as poor, regular or brilliant ones. Their simulation study is just a proof of concept for the practicality of fuzzy computation in adaptive testing. However, this study differs from their work on the fuzzification of question complexity levels, evaluation of student achievement on per competency and experimentation of the suggested mechanism on a real environment and actual educational setting, while presenting more strong indicators than that of their work for the practicality of fuzzy logic in adaptive testing.

Balas-Timar and Balas [37] have compared the efficiency of fuzzy logic computation with Bayesian likelihood estimation on question selection in adaptive tests. However, in this study the efficiency of the fuzzy logic computation is interpreted with a comparison of the results with that of classical (fixed question set) tests.

Additionally, in their recent work Sineglazov and Kusyk [38] examined the applicability of fuzzy logic computation in a cumulative test assessment setting. Their findings are consistent with that of in this study. Such that, in this study findings indicate when the test results are considered as a cumulative measurement, that is free from competencies, the difference between the classical and fuzzy-logic-based adaptive testing is not significant.

In conclusion, the contributions of the current study are summarized in the following lines.

With the findings in this experimental study, it is pointed out that the fuzzy logic values and the fuzzy logic calculation methodologies can effectively be used in question selection algorithms of computerized adaptive tests.

Furthermore, the fuzzy logic calculation methodology, if used in an interactive testing algorithm like the one proposed in this study, would be superior to the classical (fixed question set) tests in emphasizing student’s achievement level in competency base.

In defuzzification of the fuzzy assessments values, Center of Maximum Value calculation methodology is found to produce more reasonable crisp values then Center of Gravity calculation.

It is found that there is no significant difference between students’ selection of initial difficulty level and the usage of middle difficulty level at the beginning in adaptive testing, on both effectiveness of the achievement calculation and exploitation of the questions in database.

It is inferred that having priori assessment via utilization of multi-dimensional questions would rationally contribute to the measurement in a dimension. Especially, a positive priori consideration would reasonably decrease the number of questions asked to measure success level in a dimension.

Moreover, studies on backtracking drawback in adaptive testing and incorporating a self-learning mechanism for the fuzzy membership values would be valuable contributions to the automatic adaptive testing studies.

REFERENCES

[1] Hoge D. Robert and Coladarci Thedore, “Teacher Based Judgements of Academic Achievement: A Review of Literature”, *Review of Educational Research*, 59(3), 297-313, 1989.

[2] H. Borko, R. Cone, N. A. Russo, R. J. Shavelson, **Teachers’ decision making. Research on Teaching: Concepts, findings and Implications**, 136-160. Berkeley, CA: McCutchan Publishing Corporation, 1979.

[3] Penelope L. Peterson, "Teachers' and students' cognitional knowledge for classroom teaching and learning." *Educational researcher*, 17(5), 5-14, 1988.

[4] J. B. Carroll, “Measurement and Educational Psychology (Chapter 5)”, Historical Foundations of Educational Psychology, J. A. Glover, Ronning R. R. Springer - Lenum Press New York and London, 89-106, 1987.

[5] Thomas C. Reeves, "Keys to successful e-learning: Outcomes, assessment and evaluation." *Educational Technology*, 42(6), 23-29, 2002.

[6] E. L.Baker, O. F. Harold., Assessing Instructional Outcomes, U.S. Department of Education National Institute of Education Educational Resources Information Center (ERIC), Washington DC.,URL: https://files.eric.ed.gov/fulltext/ED266175.pdf, 02.01.2020.

[7] T. Evangelos, E. Georgidau, A. A. Economides, “The design and evaluation of a computerized adaptive test on mobile devices”, *Computers & Education*, 50(4), 1319–1330, 2008.

[8] H. Özcan, B. G. Emiroğlu. "Bulut Tabanlı Öğrenme Yönetim Sistemi Seçiminde Bulanık Çok Kriterli Karar Analizi Yaklaşımı." *Bilişim Teknolojileri Dergisi* 13(1), 97-111, 2020.

[9] O. Güler, O. Erdem. "Mesleki Eğitimde İnteraktif 3D Eğitimin Uygulanması ve Stereoskopik 3D Teknolojisi Kullanımı." *Bilişim Teknolojileri Dergisi* 7(3), 11.

[10] D. J. Weiss, "Improving measurement quality and efficiency with adaptive testing.", *Applied psychological measurement*, 6(4), 473-492, 1982.

[11] Ho, Rong-Guey, Yung-Chin Yen, "Design and evaluation of an XML-based platform-independent computerized adaptive testing system.", *IEEE Transactions on Education*, 48(2), 230-237, 2005.

[12] B.S. Ahmed, K. Z. Zamli, "A variable strength interaction test suites generation strategy using particle swarm optimization.", *Journal of Systems and Software*, 84(12), 2171-2185, 2011.

[13] S. Zygouris, M. Tsolaki, "Computerized cognitive testing for older adults: a review.", *American Journal of Alzheimer's Disease & Other Dementias*, 30(1), 13-28, 2015.

[14] P.J. Muñoz-Merino, M.F. Molina, M. Muñoz-Organero, C.D. Kloos, “An adaptive and innovative question-driven competition-based intelligent tutoring system for learning.”, *Expert Systems with Applications*, 39(8), 6932-6948, 2012.

[15] Internet: L.M. Rudner, “An On-Line, Interactive, Computer Adaptive Testing Tutorial”, http://EdRes.org/scripts/cat, 02.01.2020.

[16] M. Lilley, T. Barker and Carol Britton, "The development and evaluation of a software prototype for computer-adaptive testing.", *Computers & Education*, 43(1-2), 109-123, 2004.

[17] R.D. Carlson, "Computer adaptive testing: A shift in the evaluation paradigm.", *Journal of Educational Technology Systems*, 22(3), 213-224, 1994.

[18] R. L. Jacobson, “New computer technique seen producing a revolution in educational testing.”, *Chronicle of Higher Education*, 40(4),22–23, 1993.

[19] A.M. Boyd, **Strategies for controlling testlet exposure rates in computerized adaptive testing systems**, PhD Thesis, The University of Texas at Austin, May 2003.

[20] T.J.H.M. Eggen, **Overexposure and underexposure of items in computerized adaptive testing**, Measurement and Research Department Reports 2001-1, Citogroep Arnheim.

[21] A. Coşkun, R. Kılıç. "Meslek liselerinde modül değerlendirme sınavlarının çevrimiçi uygulanması”, *Bilişim Teknolojileri Dergisi* 4(1), 2011.

[22] F.M. Lord, M.R. Novick, A. Birnbaum, **Statistical theories of mental test scores**, Information Age Publishing, 2008.

[23] M.K. Sugeno, K. Asai, T. Terano, **Fuzzy Systems Theory and Its Applications**, Academic Press Limited, London, 1992.

[24] J. Ma and D. Zhou. "Fuzzy set approach to the assessment of student-centered learning.", *IEEE Transactions on Education*, 43(2), 237-241, 2000.

[25] I. Saleh, S. Kim, "A fuzzy system for evaluating students' learning achievement", *Expert systems with Applications*, 36(3), 6236-6243, 2009.

[26] S.M. Bai, S.M. Chen, "Automatically constructing grade membership functions of fuzzy rules for students' evaluation", *Expert Systems with Applications*, 35(3), 1408–1414, 2008

[27] K.Z. Zamli, F. Din, S. Baharom, B.S. Ahmed, "Fuzzy adaptive teaching learning-based optimization strategy for the problem of generating mixed strength t-way test suites", *Engineering Applications of Artificial Intelligence*, 59, 35-50, 2017.

[28] M. Badaracco, L. MartíNez, "A fuzzy linguistic algorithm for adaptive test in Intelligent Tutoring System based on competences", *Expert Systems with Applications*, 40(8), 3073-3086, 2013.

[29] J. Marciniak, "Building intelligent tutoring systems immersed in repositories of e-learning content", *Procedia Computer Science*, 35, 541-550, 2014.

[30] M. McAlpine, **A Summary of Methods of Item Analysis, Computer Assisted Assessment Center**, Blue Paper 2, b21, Luthon, 2002.

[31] Internet: L.M. Rudner, Item Response Theory (IRT), http://edres.org/irt, 02.01.2020.

[32] A. Sadollah, "Introductory chapter: which membership function is appropriate in fuzzy system?", **Fuzzy logic based in optimization methods and control systems and its applications**. IntechOpen, 2018.

[33] O. A. M. Ali, A. Y. Ali, B. S. Sumait, "Comparison between the effects of different types of membership functions on fuzzy logic controller performance", *International Journal*, 76, 76-83, 2015.

[34] J. M. Keller, D. Liu, D. B. Fogel, **Fundamentals of computational intelligence: Neural networks, fuzzy systems, and evolutionary computation**, John Wiley & Sons, 2016.

[35] S. Rajasekaran, G. A. Vijayalakshmi Pai, **Neural Networks, Fuzzy Systems and Evolutionary Algorithms: Synthesis and Applications**, PHI Learning Pvt. Ltd., 2017

[36] J. Suarez-Cansino, R. A. Hernandez-Gomez, "Adaptive testing system modeled through fuzzy logic", **2nd WSEAS Int. Conf on Computer Engineering and Applications (CEA 2008)**, Acapulco, Mexico, January. 2007.

[37] Balas-Timar, Dana V., Valentina E. Balas, "Ability estimation in CAT with fuzzy logic", **2009 4th International Symposium on Computational Intelligence and Intelligent Informatics**. IEEE, 2009.

[38] V. M. Sineglazov, A. V. Kusyk, "Adaptive testing system based on the fuzzy logic", *Electronics and control systems*, 2, 85-91, 2018.