Examination of Graders' Satisfaction with the Implementation of Open-Ended Questions in Open and Distance Learning

Assoc.Prof.Dr. Murat Akyıldız
Anadolu University, Open Education Faculty
muratakyildiz@anadolu.edu.tr
ORCID: 0000-0001-5069-0132

Assist.Prof.Dr. Nejdet Karadağ
Anadolu University, Open Education Faculty
nkaradag@anadolu.edu.tr
ORCID: 0000-0002-9826-1297

Assist.Prof.Dr. Belgin Boz Yüksekdağ
Anadolu University, Open Education Faculty
bboz@anadolu.edu.tr
ORCID: 0000-0003-2862-3544

Lecturer Ali İhsan İbileme
Eskisehir Technical University
aliihsan@anadolu.edu.tr
ORCID: 0000-0002-4705-3973

ABSTRACT
This study aims to examine graders' satisfaction with the open-ended question implementation in one of the open education systems in Türkiye. The quantitative research method was adopted in this study in which the data were collected through a questionnaire developed by the researchers. The graders' satisfaction regarding open-ended question implementation was discussed in three dimensions. These are "the functioning of the open-ended question grading system", "the need for communication/help while using the system" and "the importance of open-ended questions in open and distance learning". Whether the sub-dimension scores of the measurement tool were different from the expected values was examined with the Wilcoxon test. The graders' satisfaction regarding open-ended question implementation was analysed according to gender, title, and course variables. It was observed that the graders' satisfaction levels regarding the "Satisfaction with the grading system " and the "importance of open-ended questions in open and distance learning" were high. It was observed that the satisfaction levels of the graders regarding the " Satisfaction with the grading system", "the need for communication/help arising during using the system" and "the importance of open-ended questions in open and distance learning" did not change according to gender, academic title and the type of course they scored.

INTRODUCTION
Assessment is an important process in terms of providing an opportunity to evaluate education in terms of time, effort, and cost by determining whether learners have achieved their learning goals, and deciding whether to continue after the process, whether it is successful, whether or not to proceed to the next stage, depending on the characteristics of the learner (Nitko, 2004). In open and distance education, the process of specifying the observed qualities and observation results with numbers or symbols and making a value judgment about the measured quality by comparing the measurement results with a criterion is particularly important as it can be used to motivate distance learners who take responsibility for their learning (Harlen & Deakin-Crick, 2003). In the open and distance learning system; the process of evaluating students includes limitations compared to the traditional education environment because the instructor and learners are physically in different environments (Puspitasari, 2010). The reasons such as the fact that learners are of different ages and occupations, enrol in programs for different purposes, have various learning materials, and have different criteria for success make assessment a problematic process in open and distance education (Thorpe, 1988). In this context, instructors do not have many options to measure the performance of learners in open and distance education (McIsaac & Gunawardena, 1996). In distance education, some of the measurement and assessment activities are organized based on study materials prepared according to self-learning principles; in determining the success level of students, mostly time-limited, supervised exams are used. Today, although technological developments offer new opportunities to instructors in this sense, it is not entirely possible to ensure identity control and reliability in new learning environments. However, Tomas et al. (2015) state that distance education in higher education is increasing rapidly, and technology-based assessment is progressing more slowly than expected.

In mega universities providing open and distance education services around the world, multiple-choice tests called objective tests are used as a basic measurement tool, as well as homework and portfolio, in determining the success levels of learners. Research conducted on these institutions has revealed that learners prefer different measurement tools such as true-false tests, matching tests, homework/projects, and graduation thesis in measuring their knowledge (Karadağ, 2014). In the study of Cabı (2016), it was seen that distance learners wanted to be evaluated with exams. Because each question type has superior and non-superior aspects, it is also very important to increase the reliability of using different question types in measurement tools to turn the superiority of different question types over each other into an advantage (In'nami & Koizumi, 2009). In this context, it is not possible to determine whether high-level thinking skills are acquired or not with tests in which only multiple-choice questions are used (Husain et al., 2012).

Open-ended questions are the most appropriate question types used to measure high-level skills such as problem-solving, organizing problems, generating new and original ideas, evaluating ideas, applying the information in different situations, establishing cause-effect relationships, making generalizations, generating hypotheses, making comparisons between alternatives (Kwon et al., 2006; Foong, 2002). Long mixed tests are moderately difficult and more distinctive than tests consisting of only multiple choice and only open-ended items (Kurniawan et al., 2018). The tests, which include open-ended questions, focus on perception, justification, and the ability to use information, and that such questions allow learners to reflect on their differences (Melovitz Vasan et al., 2017; Wooten et al., 2014). On the other hand, the answers given by the learners to the open-ended questions can also be used to obtain information about the quality of the learning process (Lee, Liu, & Linn, 2011). It is stated that open-ended questions expressing classical, written questions that cannot be answered simply as yes or no and that do not offer options are the most appropriate question type to measure high-level skills. At the same time, other superior aspects of open-ended questions are that they reduce the measurement error by eliminating chance success, are suitable for partial scoring, and can be prepared more easily than multiple-choice items (Allan, & Driscoll 2014; Ventouras et al., 2011). However, in the use of open-ended questions; There may be situations such as taking a long time to implement and scoring, difficulty in providing content validity, fewer questions due to time constraints, and most importantly, inability to score objectively (Palmer, & Devitt, 2007; Reiner et al., 2003).

When the studies on evaluation with open-ended questions in universities that implement open and distance education in the world are examined, it is seen that the evaluation made with open-ended questions from both learners and instructors is subjective (Aisha, 2007). For example, in AIOU, the success of students is determined by homework and exams with three hours of open-ended questions. Scoring is done by the teaching staff working at universities throughout the country, and when these are not sufficient, with the contribution of teachers affiliated to the Ministry of National Education. An automation system is not used and the announcement of the exams takes a long time like one month. In this situation, when learners are informed about the process and how the scoring will be, students' anxiety about grader reliability can be reduced.

The difficulty of distinguishing between true and false in open-ended questions would make it difficult to read and score the paper. On the other hand, the majority of the educators scored the questions according to the degree of difficulty, that they gave high points to difficult questions and lower points to easy questions, and that they did not prefer to give equal points to each question. At the same time, the student's inability to remember information during the exam, insufficient writing skills, or prejudices about the difficulty of open-ended questions are also conditions that affect student performance and the use of open-ended questions (Reiner et al., 2003).

An under-graduate level open education system in Türkiye has also started to determine student success with tests that include open-ended questions in the 2017-2018 academic year fall semester exams. During this period, two short-answer questions worth 5 points, one long-answer question worth 10 points, and 16 multiple-choice questions worth 5 points were included in the tests consisting of 19 questions in certain courses. With the long answer question, learners are expected to express their views and thoughts on the given topic by writing a paragraph. Open-ended questions are evaluated with the Open-Ended Question Grading (OEQG) System designed online. Until the open-ended question implementation, which is the subject of this study, it is the first example in the context of open and distance learning throughout the country. The details of the installed system are given below.

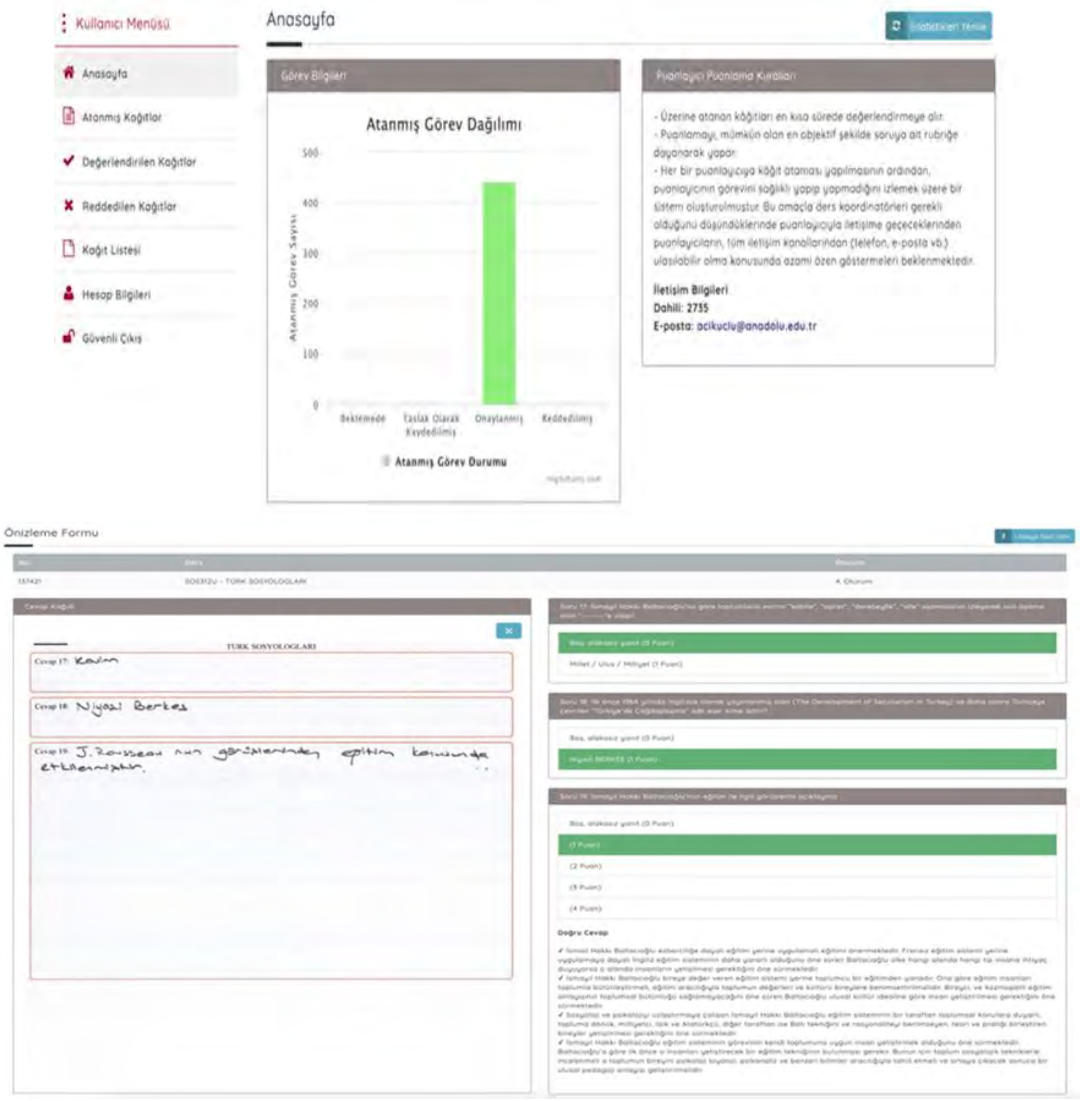
Open-Ended Question Grading (OEQG) System

Within the scope of Open Education System (OES), open-ended questions are asked to students in midterm exams in six different departments at the undergraduate level. In this context, an infrastructure has been prepared by the Computer Research and Implementation Centre (CRIC) Web Group for the grading of open-ended questions. The open-ended question system consists of two different stages. The first stage is the implementation of the open-ended exam, and the second stage is the grading of the open-ended exam. During the implementation phase of the

open-ended exam, the long and short-answer questions that are planned to be asked in the exam are taken from the course editor by the Test Research Unit. The determined questions are presented to the students after the multiple-choice questions in the question booklet containing the related courses. In an exam, a total of three open-ended questions are asked, two with short answers and one with long answers, for a course. In addition to the optical form prepared for multiple choice questions, a separate answer sheet is prepared for the students who take the course in the program where open-ended questions will be asked by the Computer Research and Implementation Centre Exam Software team. An area is left on this answer sheet where students can enter long answer questions along with fill-in-the-blank or short answer questions. While students enter the answers to the multiple-choice questions in the optical form, they enter their answers to the open-ended questions on the answer sheet. After the exam implementation phase, the answer sheets that come to CRIC are scanned by scanner devices and converted into digital form. After this stage, the scanned papers are sent to the OEQG system and the grading process begins. There are four different user roles in OEQG: Observer, Section Coordinator, Referee, and Grader. The duties of these users in the OEQG system are described below.

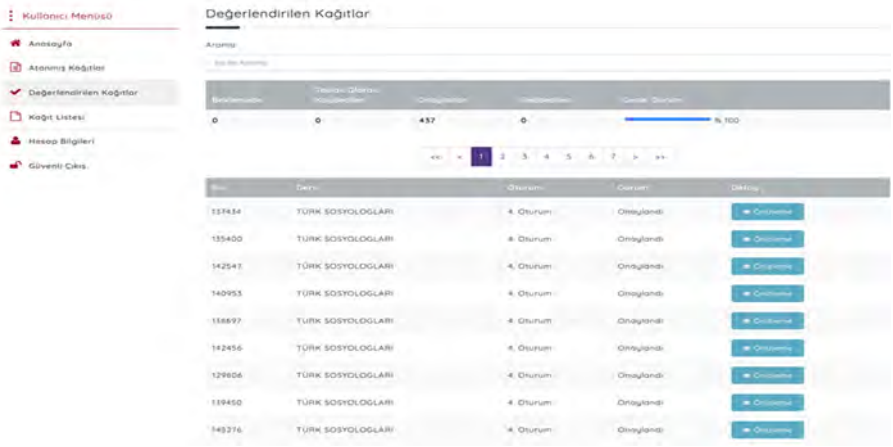
Grader Role: They are the users who evaluate the papers automatically assigned by the system according to certain criteria.

Figure 1. Grader Page View



Referee Role: In case of conflict in the papers evaluated by the graders, they are the reviewers.

Figure 2. Referee Page View

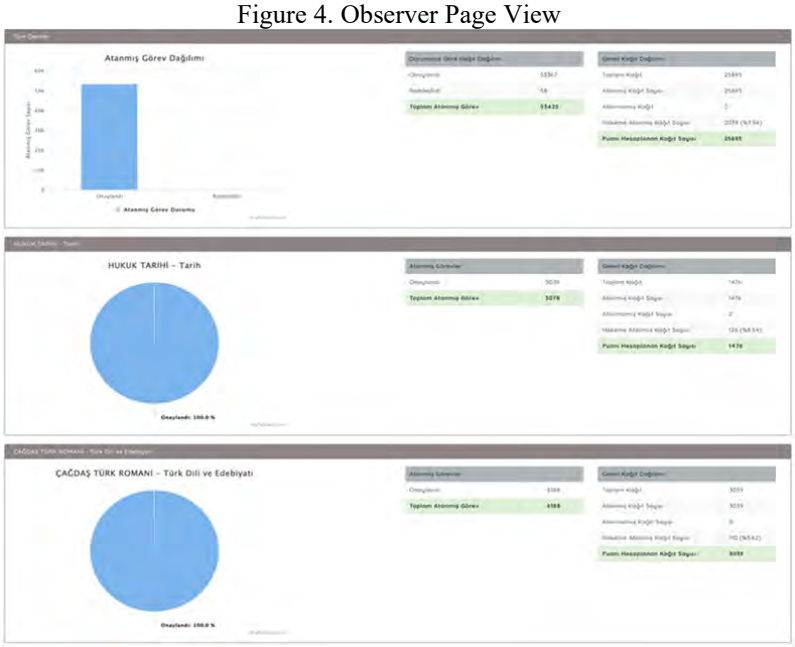


Program Coordinator Role: Each program has a coordinator. The user follows the referees and scorers in the program. Monitoring of scorers, change of scorers, tracking of rejected papers and assignment to referees, etc. performing transactions.

Figure 3. Program Coordinator Page View

Observer Role: Users who monitor the entire system over the OEQG system.

Figure 4. Observer Page View



OEQG Workflow Process

Papers scanned over the OEQG system are uploaded to the system at regular intervals. After this loading process, each paper is automatically assigned to two different graders responsible for the course. After each assignment, the graders are informed via SMS and e-mail. At the same time, within the period determined during the grading phase, task reminders are made to the graders at regular intervals every day. Until all the papers belonging to the students are scanned, paper assignments continue to be made to the graders at regular intervals. The path followed during the grading of a paper is explained below. Papers assigned by each grader are evaluated based on the rubric. For long answer questions, a rubric with a minimum of three or more options is presented. In short answer questions, a rubric consisting of true/false options is presented. All papers are shown to the scorers before the assignment. In this way, the rubrics are organized by examining the answers given by the students by the graders. In this direction, it has been tried to reduce the risk of error in the grading by considering all possibilities. Different graders evaluate each paper to make a fair grading. If there is a conflict between the points given by the scorers, the paper is sent to the referee.

The referee can only take action on the disputed question. For example, if the first grader gives two points and the second grader gives four points in a long answer question on a paper, a conflict arises. In this case, the referee only sees the long-answer question and can act on it. After the referee evaluates the disputed question, the average score of the paper is calculated by the system according to the weight of the questions in the rubric. If there is no disagreement, the average score of the paper is calculated according to the weight of the questions by taking the average of the scores given by the graders.

Determining the satisfaction levels regarding open-ended questions, which have an important place in the grading activities of learners in increasing the quality in open and distance learning environments, is very important in terms of improving the practice. In this study, although it changes in each period, how the answers of 181,162 learners to open-ended questions in their own handwriting are perceived by 1411 raters, 58 referees, 4 coordinators and 4 observers, who evaluated their answers in 72 hours, were examined from various perspectives. In this context, the purpose of this research is to examine the satisfaction of graders with open-ended questions in the Open Education System. The graders' satisfaction with open-ended questions was discussed in three dimensions. These are "Satisfaction with the grading system", "The need for communication/assistance during using the system" and "The importance of open-ended questions in open and distance learning".

THE STUDY

This study uses quantitative research method. Design of the study is survey. Data were gathered by a scale developed by researchers. In this study, answers to the following questions were sought:

RQ1: Do graders' scores on subtests differ from the midpoint subtest score level?
RQ2: Do graders' sub-dimension scores differ according to gender?
RQ3: Do graders' sub-dimension scores differ according to the courses they grade?
RQ4: Do graders' sub-dimension scores differ according to their academic titles?

Participants

The participants of the research are 169 instructors who work as graders in the open-ended question implementation in OES in the 2018-2019 academic year. The distribution of the graders participating in the study regarding their gender and academic titles is given in Table 1.

Table 1. Distribution of Graders By Gender And Academic Titles

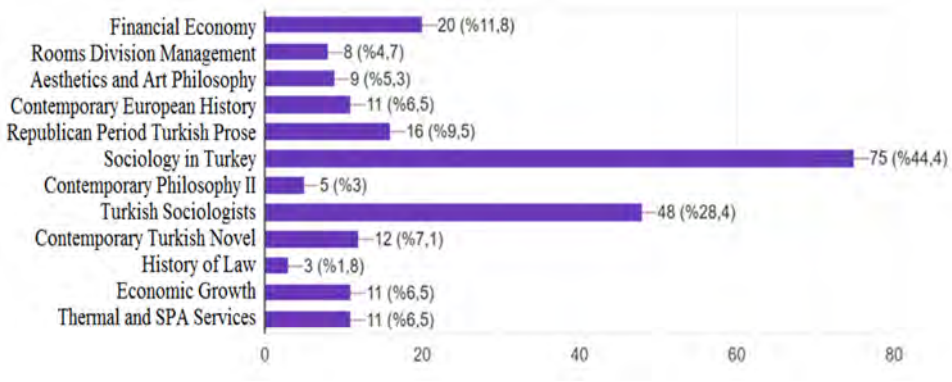
Gender	**N**	**%**
Female	98	58
Male	71	42
Academic Title		
Professor doctor	13	7.7
Associate professor	29	17.2
Assistant professor	28	16.6

Lecturer doctor	8	4.7
Lecturer	26	15.4
Research assistant doctor	9	5.3
Research assistant	56	33.1
Total	169	100

Table 1 shows that while women constitute 58% of the graders participating in the research, the rate of men is 42%. Looking at the distribution of graders by academic titles, the title with the highest percentage is Research Assistant (33.1%).

The distribution of the courses that the participants worked on is given in the chart below.

Chart 1. The courses with open-ended question practice



As seen in Chart 1, the course with the highest number of graders was "Sociology in Türkiye" (44.4%). This is followed by the Turkish Sociologists (28.4%) course. These two courses are taught in the Sociology Undergraduate Program, where the number of students in the Open Education System is high. The courses with the lowest number of graders are History of Law (1.8%) and Contemporary Philosophy-II (3%). This shows that the number of graders is determined in direct proportion to the number of students taking the course.

Data Collection

In the research, the data were collected with the help of a measurement tool prepared via Google Forms. In the measurement tool, there are 16 closed-ended and 1 open-ended questions, 11 of which are Likert type and 5 of which are answered as yes-no. In the open-ended question, graders were asked to state their opinions on other issues that were not included in the questionnaire regarding the implementation.

Since the items of the measurement tool used in the study were developed in line with 3 purposes, confirmatory factor analysis was applied to determine the extent to which the items measure these 3 objectives. The first factor, in which the items are grouped, is the satisfaction with the grading system, the second factor is the need for communication/help while using the system, and the third factor is the perceived importance of open-ended questions in open and distance learning. Since factor scores were thought to be correlated with each other, analyses were performed by allowing correlations between factors. Since the items were ordinally ranked, confirmatory factor analysis was applied using the diagonally-weighted least squares method (DWLS) and oblimin rotation.

As a result of the confirmatory factor analysis, Chi-Square value was 153.796, degrees of freedom 101, RMSEA value was 0.056, SRMR value was 0.110, CFI value was 0.95, NFI value was 0.87, TLI value was 0.94. When these fit indices were examined, it was seen that the 3-factor model has a good fit with the data (Kenny, 2020; Shi et al., 2019, Kline, 2016). The single-factor construction of the data was considered as an alternative model and tested against the 3-factor model. The RMSE value of this model was 0.081, the CFI value was 0.89, and the TLI value was 0.873. The difference between the chi-square values of the 3-factor model and the single-factor model (41.276, sd difference= 3) was significant in favor of the 3-factor model. The difference between the CFI values of the two models is larger than the 0.01 as suggested by Cheung and Rensvold (2002), in favor of the three-factor model. In line with these findings, it was decided that the scale measures three correlated dimensions: satisfaction

with the open-ended question system, the need for communication/help while using the system, and opinions on asking open-ended questions. Correlations between subscales were -0.326 between the first and second factors; 0.597 between the first factor and the third factor; It was observed as -0.140 between the second factor and the third factor. All correlations are statistically significant at the 0.05 level. The omega reliability coefficients of factors were respectively 0.74, 0.58 and 0.52.

All 169 instructors who participated in the open-ended question scoring process were contacted via e-mail and SMS, and the survey link and information letter were sent. The contribution of the study to the improvement of the system was mentioned in the information letter and it was stated that the participation was on a voluntary basis. The answers given to the measurement tool were collected between 14-20 February 2020.

Analysis of Data

Before the analyses data were checked in terms of normality and outliers. 14 outliers in the 2nd factor and 6 outliers were detected in the 3rd factor scores. All analyses were repeated by removing these extreme values from the analyses and by not removing them, and it was decided to use complete data as it was seen that extreme values did not change the results. Data were not normally distributed in most of the subgroups before and after removal of outliers. Normality tests were given before the corresponding analysis below.

Since the subscale scores were not normally distributed, Wilcoxon signed test was used to find an answer to the 1st question. Mann-Whitney test was used to find an answer to the 2nd and 3rd questions. Kruskal-Wallis test was used to find an answer to the 4th questions of the study.

In general, item level analysis results are not necessary but in order to inform readers about items, chi-square independence test was applied to each item to determine whether there is a significant independence between the agreeing/disagreeing level and gender, course type and academic title.

FINDINGS

In order to answer the first question, the midpoint scores (midpoint of possible min and max scores of subscales) of subscales were used as average satisfaction, need and perceived importance levels. Midpoints as average values frequently used as a cut-off point (Zheng et al., 2011; McDonald et al., 2014). Each subscale has its own score range and midpoint. Since subscale scores were not normally distributed (Shapiro-Wilkfactor1=0.937, n=169; p<0.05; Shapiro-Wilkfactor2=0.757, n=169, p<0.05, Shapiro-Wilkfactor3=0.968, n=169; p<0.05), one sample Wilcoxon signed test was used to test whether median subscale levels differs than average subscale value (Table 2).

Table 2. Wilcoxon Test Results

	Value	Sig.	Rank-biserial r
Satisfaction with the grading system **Average satisfaction level =15, $\bar{X}$=33.207, median=33 (min=24, max=37)**	14365.000	< .001*	1.00
Need for communication/assistance while using the system **Average need level =1.5, $\bar{X}$=0.757, median=0 (min=0, max=3)**	2268.000	< .001*	- 0.684
The importance of open-ended questions **Average importance level =12, $\bar{X}$=14.053, median=14 (min=7, max=20)**	10314.500	< .001*	0.822

*p<0,01

When Table 2 is examined, it is seen that the participant's satisfaction with the "grading system" (median=33) is significantly higher than the average satisfaction level (midpoint = 15; W=14365, p<0,05). The communication/help needs of the participants (median = 0) while using the system were significantly lower than the median need level of 1.5 (midpoint = 1.5; W=2268). The participants' median scores (12) on the perceived importance of open-ended questions in open and distance learning were significantly higher than the average perceived importance level (midpoint = 12; W=10314.5). All comparisons have significantly high rank biserial correlation-based effect sizes.

In order to answer the second question, score distributions of gender groups were examined and not found normal. Normality test results and descriptive statistics given below in Table 3.

Table 3. Normality Tests and Descriptive Statistics Table For Gender

	Group	N	Median	Shapiro-Wilk	p	Min	Max
Satisfaction with the grading system	Women	98	34	0.927	< .001	24	37
	Men	71	33	0.94	0.002	25	37
Need for communication/assistance while using the system	Women	98	1	0.774	< .001	0	3
	Men	71	0	0.725	< .001	0	3
Importance of open-ended questions	Women	98	14	0.964	0.009	8	18
	Men	71	14	0.95	0.006	7	20

As can be seen in Table 3 none of the distributions were normal. In order to test whether the mean rank values of subscale scores differ significantly across gender groups, Mann-Whitney test was applied. The results are given in Table 4.

Table 4. Mann-Whitney Test Results for Gender

Subscale	Group	Mean rank	W	P
Satisfaction with the grading system	Women	87.13	3270	0.504
	Men	82.06		
Need for communication/assistance while using the system	Women	89.09	3078.5	0.164
	Men	79.36		
Importance of open-ended questions	Women	89.23	3779.5	0.335
	Men	81.94		

Mann-Whitney test indicated that satisfaction with the grading system, need for communication while using the system and perceived importance of open-ended questions was not significantly higher in any gender. The results indicated that graders from each gender have the same satisfaction with the grading system, require the same need for communication while grading and give the same importance to the open-ended questions. As previously shown, graders' median scores were higher than the average score in each sub dimension. Therefore, it can be concluded that graders of both genders have high scores in each sub dimension.

In order to answer the third question, score distributions of course groups were examined and not found normal. Normality test results and descriptive statistics given below in Table 5.

Table 5. Normality Tests and Descriptive Statistics Table For Courses

Subscale	Group	N	Median	Shapiro-Wilk	p	Min	Max
Satisfaction with the grading system	Numeric	20	33	0.910	0.065	25	37
	Verbal	149	34	0.939	< .001	24	37
Need for communication / assistance while using the system	Numeric	20	0	0.626	< .001	0	1
	Verbal	149	1	0.767	< .001	0	3
Importance of open-ended questions	Numeric	20	13	0.930	0.152	8	20
	Verbal	149	14	0.966	0.006	7	20

Normality tests showed that only numeric class graders' scores are normal in the first and the last dimension. All other groups' score distributions are not normal. In order to test whether the mean rank values of graders' subscale scores differ significantly across verbal and numeric course type Mann-Whitney test was applied. The results are given in Table 6.

Table 6. Mann-Whitney Test Results for Course Type

Subscale	Group	Mean rank	W	p
Satisfaction with the grading system	Numeric	80.73	1404.5	0.677
	Verbal	85.57		
Need for communication/assistance while using the system	Numeric	71.10	1212	0.140
	Verbal	86.87		
Importance of open-ended questions	Numeric	54.10	872	0.002
	Verbal	89.15		

Mann-Whitney test indicated that mean rank values of satisfaction with the grading system and need for communication while using the system was not significantly higher in any course type. On the contrary there is a significant difference between mean rank values of grading course type ($W=872$, $p<0.05$). Graders' who were grading numeric courses have higher mean rank values than those grading verbal courses which means verbal course graders give more importance to open ended questions than numeric course graders.

In order to answer the fourth question, score distributions of groups based on academic title were examined and not found normal. Normality test results and descriptive statistics given below in Table 7.

Table 7. Normality Tests and Descriptive Statistics Table For Courses

Subscale	Group	N	Median	Shapiro-Wilk	p	Min	Max
Satisfaction with the grading system	Lecturer	26	34	0.915	0.035	28	37
	Res. Asst.	56	33	0.935	0.003	26	37
	Asst. Prof	45	34	0.944	0.029	27	37
	Assoc. Prof	29	33	0.925	0.041	25	37
	Prof	13	30	0.911	0.188	24	37
Need for communication / assistance while using the system	Lecturer	26	0	0.807	< .001	0	3
	Res. Asst.	56	0	0.785	< .001	0	3
	Asst. Prof	45	0	0.751	< .001	0	3
	Assoc. Prof	29	0	0.632	< .001	0	2
	Prof	13	0	0.662	< .001	0	3
Importance of open-ended questions	Lecturer	26	8	0.883	0.007	8	17
	Res. Asst.	56	9	0.920	0.001	9	20
	Asst. Prof	45	8	0.958	0.104	8	20
	Assoc. Prof	29	7	0.955	0.242	7	18
	Prof	13	11	0.946	0.536	11	18

Normality tests showed that Prof's scores are normally distributed in the first and the last dimension. In the last dimension Asst. Prof, Assoc. Prof and Prof groups have normal distributions. All other conditions have non-normal distribution. Therefore, Kruskal-Wallis test was used to compare mean ranks of subgroups in each dimension. The results are given in Table 8.

Table 8. Kruskal-Wallis Test Results for Graders' Title

Subscale	Group	Mean rank	H	Df	p
Satisfaction with the grading system	Lecturer	92.79	2.528	4	0.640
	Res. Asst.	87.62			
	Asst. Prof	81.43			
	Assoc. Prof	85.78			
	Prof	68.77			
Need for communication/assistance while using the system	Lecturer	91.13	5.708	4	0.222
	Res. Asst.	92.26			
	Asst. Prof	84.40			
	Assoc. Prof	69.67			
	Prof	77.73			
Importance of open-ended questions	Lecturer	77.94	3.455	4	0.485
	Res. Asst.	90.63			
	Asst. Prof	90.68			
	Assoc. Prof	75.48			
	Prof	76.46			

Kruskal-Wallis test indicated that mean rank values of graders who had different academic titles had no significant difference for all of the subscales. In other words, each group which has a different academic title has the same mean rank score in each subscale. Therefore, it seems that the academic title of graders doesn't change the mean rank values of the scores. Based on these results all graders from different academic levels have the same level of acceptance for open-ended questions and satisfaction with the grading system.

As mentioned before item level analysis results are not necessary but in order to inform readers about items, chi-square independence test was applied to all items using gender, course type and academic title. Likert type item structure recoded as agree and not agree. For this purpose, item scores including 1, 2 and 3 were coded as disagree, 4 and 5 were recoded as agree. Gender, course type and academic title were used as is. Results are given below in Table 9.

Table 9. Item Level Chi-Square Independence Test

Statements	Gender	Academic Title	Type of Class
	X2 (df-p)	**X2 (df-p)**	**X2 (df-p)**
Before I started scoring open-ended questions, I had enough information about the scoring process.	1.02 (4-0.600)	6.55 (36-0.886)	1.26 (4-0.533)
I find it necessary to implement open-	0.287 (4-0.991)	36.6 (36-0.048)*	8.81 (4-0.066)

ended questions.			
I think that students get more points from open-ended questions than they deserve.	6.31 (4-0.177)	23.8 (36-0.476)	6.54 (4-0.162)
I think that open-ended questions measure students' knowledge better than multiple-choice questions.	6.05 (4-0.196)	22.6 (36-0.544)	6.48 (4-0.166)
In the answers of the students, I encountered original/different answers that could be added to the answer key.	1.15 (4-0.885)	16.1 (36-0.883)	10.6 (4-0.03)*
I think that saving as a draft in scoring contributes positively to the scoring process.	5.27 (4-0.261)	26.0 (36-0.355)	5.60 (4-0.231)
I could easily access the papers assigned in the open-ended question scoring software.	5.14 (4-0.162)	28.0 (36-0.063)	5.25 (4-0.154)
I was able to evaluate the questions easily in the open-ended question scoring software.	3.53 (4-0.316)	10.4 (36-0.918)	0.889 (4-0.828)
I think the paper assignment process is done in appropriate periods.	2.93 (4-0.570)	17.5 (36-0.827)	2.71 (4-0.607)
I found the open-ended question scoring software useful.	4.47 (4-0.346)	24.9 (36-0.411)	0.742 (4-0.946)
With the open-ended question scoring software, I was able to follow the system instantly through informative SMS/e-mail throughout the process.	0.824 (4-0.672)	15.9 (36-0.198)	0.512 (4-0.774)
The time given for scoring the open-ended questions was sufficient.	0.01 (1-0.993)	16.9 (6-0.010) *	0.071 (1-0.790)
Did you feel the need to communicate with other graders while scoring open-ended questions?	1.20 (1-0.273)	12.0 (6-0.061)	3.25 (1-0.07)
Did you feel the need to communicate with the referee while scoring open-ended questions?	0.352 (1-0.553)	4.13 (6-0.659)	0.014 (1-0.906)
Did you feel the need to communicate with the coordinator while scoring open-ended questions?	1.27 (1-0.260)	3.13 (6-0.793)	3.33 (1-0.068)
Did you find the open-ended question grading guide video helpful?	0.087 (1-0.768)	10.10 (6-0.12)	2.76 (1-0.096)

*p<0.05

As can be seen in Table 9, three of the items have significant dependence. Most of the items has no significant dependency with gender, course type or academic title.

CONCLUSIONS

When the findings were evaluated, it was seen that the need for communication/help while using the system were

lower than the average need. This result can be attributed to graders not needing help before, during, and after the scoring process. In this context, it can be said that the OEQG system responds to the needs of the graders. On the other hand, it was observed that the graders' satisfaction levels were high regarding the satisfaction with the grading system and the importance of open-ended questions in open and distance learning. Alkan (2013) stated that intelligibility is provided more easily with open-ended questions, they focus on their ability to use logic and knowledge, and their success levels are revealed better than multiple-choice questions. On the other hand, since the answers to open-ended questions will show individual differences due to their perspectives, scoring methods and reliability of scoring are very important. In the related literature, the most important problem in the use of open-ended items is the inability to score objectively (Romagnano, 2001).

Each learner's answer sheet was evaluated by at least two graders to provide objective scoring in the OEQG system, which is the subject of this study. If the difference between the two scores is large, the answer sheet was sent to a referee for further evaluation. Ebel (1951) stated that the use of more than one grader in the grading process of tests consisting of open-ended questions and the means of the scores given by different graders are necessary to obtain reliable results regarding the success levels of learners. Güler et al. (2015) stated that correlations between graders greater than 0.70 reflect a high level and positive relationship. Turgut & Baykul (2012) stated that at least two and at most five people should be assigned as graders in open-ended questions, and increasing the number of graders would not provide a significant increase in the reliability of the scoring.

Dubrovich (2002) stated that there should be consistency among the graders and that the graders should be trained in scoring beforehand to ensure reliability. He emphasized that the preparation of scoring instructions (rubrics, checklists, etc.) for the feature to be measured to increase reliability would increase the consistency among the graders. In this study, rubrics for each lesson were prepared and a reliable scoring was ensured. All papers are shown to the scorers before the assignment. In this way, the rubrics are organized by examining the answers given by the students by the graders. On the other hand, studies are showing that rubrics are insufficient to eliminate grader effects such as grader strictness/generosity (Alharby, 2006). In addition to the scorer effect, the differences in course types may also be an important factor in the emergence of this situation. Interaction between grader and criterion may cause inconsistency between graders (Bıkmaz Bilgen & Doğan, 2017).

It has been emphasized in some studies that graders can be objective or biased (Köse et al., 2016; Yüzak et al., 2015). Çetin (2019) stated that although the scoring criteria were determined before the scoring process, there were differences in the scoring process among the graders, and suggested that the Multi-Faceted Rasch measurement model be used effectively in the exam gradings in higher education to solve this problem. The graders' satisfaction with each sub-dimension does not differ according to gender, academic titles, and the type of course they grade. In other words, the satisfaction levels of the graders are similar in terms of gender, course type, and titles. The participants of the study suggest that open-ended questions are important in distance education and that the implementation should be expanded with more courses. Providing the necessary training to the coordinators, graders, referees, and observers who will take part in the dissemination process is seen as one of the important activities in increasing the efficiency of the implementation.

On the other hand, it is thought that it is useful to employ different measurement tools such as open-ended questions, homework / projects, portfolios in order to measure the achievements of open and distance learners at the upper cognitive level included in the Bloom's Taxonomy. The fact that decision makers in open and distance learning focus on this issue will increase the quality of assessment practices.

It is considered that it would be beneficial to design and implement similar systems, which are the subject of this study, so that education is not interrupted by measuring high-level achievements in extraordinary situations such as COVID-19 pandemic, energy crisis, natural disasters.

REFERENCES

Aisha, R. (2007). *Evaluation of continuous assessment and final examination in teacher training programs of AIOU*. (Publication No: 1676710025670). [Master's thesis, Allama Iqbal Open University].

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs analytic, using two measurement models, the generalizability theory and the many-facet rasch measurement, within the context of performance assessment*. [Doctoral dissertation, The Pennsylvania State University].

Alkan, M. (2013). *Comparison of different designs in scoring of PISA 2009 reading open-ended items according to generalizability theory*. (Publication No: 321957) [Doctoral dissertation, Hacettepe University].

Allan, E., & Driscoll, D. (2014). The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. *Assessing Writing, 21* (7), 37-55. http://dx.doi.org/10.1016/j.asw.2014.03.001

Bıkmaz Bilgen, O., & Doğan, N. (2017). The comparison of interrater reliability estimating techniques. *Journal of Measurement and Evaluation in Education and Psychology, 8* (1), 63-78. https://doi.org/10.21031/epod.294847

Cabı, E. (2016). The perception of students on e-assessment in distance education. *Journal of Higher Education and Science, 6* (1), 94-101. https://dergipark.org.tr/en/download/article-file/1711633

Çetin, Ş. (2019). Analysis of open-ended questions with many facet rasch model. In V. Karaca (Ed), 4th International Symposium on Innovative Approaches in Health and Sports Sciences. *SETSCI Conference Proceedings*, (pp. 108-110). SETSCI. https://doi.org/10.36287/setsci.4.9.067

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9* (2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Dubrovich, M. A. (2002). Student achievement data: Holding teachers accountable. *Principal, 81* (4), 30-32, 34.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16* (4), 407-424. http://dx.doi.org/10.1007/BF02288803

Foong, P. Y. (2002). The Role of problems to enhance pedagogical practices in the Singapore mathematics classroom. *The Mathematics Educator, 6* (2), 15-31. http://hdl.handle.net/10497/52

Güler, N., & Taşdelen Teker, G. (2015). The evaluation of rater reliability of open-ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology, 6* (1), 12-24. https://doi.org/10.21031/epod.63041

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: principles, policy & practice, 10* (2), 169-207. https://doi.org/10.1080/0969594032000121270

Husain, H., Bais, B., Hussain, A., & Samad, S. A. (2012). How to construct open-ended questions. *Procedia-Social and Behavioral Sciences, 60*, 456-462. https://doi.org/10.1016/j.sbspro.2012.09.406

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: focus on multiple-choice and open-ended formats. *Language Testing, 26* (2), 219-244. https://doi.org/10.1177/0265532208101006

Karadağ, N. (2014). *Açık ve uzaktan eğitimde ölçme ve değerlendirme: Mega üniversitelerdeki uygulamalar. [Assessment in open and distance education: Practices in mega universities.*]. (Publication No: 363040) [Doctoral dissertation, Anadolu University].

Kenny, D. A. (2020). *Measuring model fit*. http://www.davidakenny.net/cm/fit.htm

Kline, R. B. (2016). *Principles and practice of structural equation modelling* (4th ed.). Guilford publications.

Köse, İ. A., Usta, H. G., & Yandı., A. (2016). Evaluation of presentation skills by using many facets rasch model. *Bolu Abant İzzet Baysal University Journal of Faculty of Education*, 16(4), 1853-1864. https://dergipark.org.tr/en/pub/aibuefd/issue/28550/304600

Kurniawan, H., Putri, R. I. I., & Hartono, Y. (2018). Developing open-ended questions for surface area and volume of beam. *Journal on Mathematics Education, 9* (1), 157-168. https://files.eric.ed.gov/fulltext/EJ1173654.pdf

Kwon, O. N., Park, J. H., & Park, J. S. (2006). Cultivating divergent thinking in mathematics through an open-ended approach. *Asia Pacific Education Review, 7*, 51-61. https://doi.org/10.1007/BF03036784

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24* (2), 115-136. https://doi.org/10.1080/08957347.2011.554604

McDonald, H., Karg, A. J., & Leckie, C. (2014). Predicting which season ticket holders will renew and which will not. *European Sport Management Quarterly, 14* (5), 503-520. https://doi.org/10.1080/16184742.2014.944199

McIsaac, M. S., & Gunawardena, C. N. (1996). *Distance education*. In D. H. Jonassen (Ed.), Handbook of research for educational communications and technology (pp. 403-437). Simon & Schuster Macmillan. http://members.aect.org/edtech/ed1/13/index.html

Melovitz Vasan, C. A., DeFouw, D. O., Holland, B. K., & Vasan, N. S. (2017). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomical sciences education, 11* (3), 254-261. https://doi.org/10.1002/ase.1739

Nitko, A. J. (2004). *Performance, portfolio, and authentic assessments: An Overview*. Educational Assessment of Students. Pearson.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple-choice questions? Research paper. *BMC medical education, 7*, 49. https://doi.org/10.1186/1472-6920-7-49

Puspitasari, K. A. (2010). Student assessment. Policy and Practice. In T. Belawati & J. Baggaley (Eds.). *Asian Distance Education* (pp. 60-65). Sage.

Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *Preparing effective essay questions*. New Forums Press. http://www.ndcl.info/Writing%20Effective%20Essay%20Questions.pdf

Romagnano, L. (2001). Implementing the assessment standards: The myth of objectivity in mathematics assessment. *Mathematics Teacher, 94* (1), 31-37. https://doi.org/10.5951/MT.94.1.0031

Shi, D., Lee, T., & Maydeu Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79* (2), 310–334. https://doi.org/10.1177/0013164418783530

Thorpe, M. (1998). Assessment and 'third generation'distance education. *Distance Education, 19* (2), 265-286. https://doi.org/10.1080/0158791980190206

Tomas, C., Borg, M., & McNeil, J. (2015). E-assessment: Institutional development strategies and the assessment life cycle. *British Journal of Educational Technology, 46* (3), 588-596. https://doi.org/10.1111/bjet.12153

Turgut, M. F., & Baykul, Y. (2012). Eğitimde ölçme ve değerlendirme (Measurement and evaluation in education). Pegem.

Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2011). Comparison of oral examination and electronic examination using paired multiple-choice questions. *Computers & Education, 56* (3), 616-624. https://doi.org/10.1016/j.compedu.2010.10.003

Wooten, M. M., Cool, A. M., Prather, E. E., & Tanner, K. D. (2014). Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory. *Physical Review Special Topics-Physics Education Research, 10* (2), 020103. https://doi.org/10.1103/PhysRevSTPER.10.020103

Yüzüak, A. V., Yüzüak, B., & Kaptan, F. (2015). A many-facet rasch measurement approach to analyze peer and teacher assessment for authentic assessment task. *Journal of Measurement and Evaluation in Education and Psychology, 6* (1), 1-11. https://doi.org/10.21031/epod.57425

Zheng, Y., Yin, T., Dong, D., & Fu, S. (2011). Using NASA-TLX to evaluate the flight deck design in Design Phase of Aircraft. *Procedia Eng. 17*, 77–83. https://doi.org/10.1016/j.proeng.2011.10.010