NOVITAS-ROYAL
RESEARCH ON YOUTH AND LANGUAGE

2020, 14(2), 76-96
Novitas-ROYAL (Research on Youth and Language)
ISSN: 1307–4733

Research Article

Analysis of Multiple-choice versus Open-ended Questions in Language Tests According to Different Cognitive Domain Levels

Murat POLAT[1]

[1]Ph.D., Anadolu University, Eskişehir, TURKEY
mpolat@anadolu.edu.tr
ORCID: 0000-0001-5851-2322

Article information

Submission	30/04/2020	Revision received	12/08/2020
Acceptance	01/09/2020	Publication date	20/10/2020

Keywords: Foreign language assessment, reliability, validity, question types, test types

Abstract: Classroom practices, materials and teaching methods in language classes have changed a lot in the last decades and continue to evolve; however, the commonly used techniques to test students' foreign language skills have not changed much regardless of the recent awareness in Bloom's taxonomy. Testing units at schools rely mostly on multiple choice questions (MCQ) because these types of questions are reliable, cost effective and time savers; however, these measure only surface information in a particular skill while other skills such as critical thinking and synthesis cannot be evaluated using MCQs. On the other hand, open-ended question (OEQ) tests include analysis and synthesis and a higher level of cognitive processing, with some washback effects such as less reliable results and more time and effort in scoring. This study aims to compare language learners' performances on MCQ and OEQ tests administered to 116 students studying at a language preparatory school in Eskisehir. Four separate tests including grammar and reading questions in both forms were prepared and administered to the same students respectively. Research results showed that there was a significant difference between OEQ and MCQ tests in terms of item difficulty and item discrimination levels. In both grammar and reading assessment, MCQ tests were found to be easier than OEQ tests.

Anahtar Sözcükler: Yabancı dilde ölçme, güvenirlik, geçerlik, soru çeşitleri, test çeşitleri

Yabancı Dil Ölçümünde Açık Uçlu ve Çoktan Seçmeli Soruların Farklı Biliş Seviyelerinde Karşılaştırılması

Özet: Yabancı dil sınıflarında görülen ders işleme teknikleri, kullanılan materyaller ve sınıf aktiviteleri son yıllarda oldukça değişmiş ve gelişmiştir, ancak Bloom'un taksonomisinin yaygın kullanımına rağmen öğrencilerin dil becerilerinin ölçümünde kullanılan sınav metotları pek fazla değişikliğe uğramamıştır. Okulların ölçme değerlendirme birimleri birtakım nedenlerden ötürü (güvenirlik, ekonomiklik, zaman tasarrufu) eleştirel düşünce, sentez kabiliyeti gibi becerileri ölçmeyip yüzeysel bilgiyi ölçen çoktan seçmeli sınavların sonuçlarına güvenmektedirler. Öte taraftan, açık uçlu sorular da analiz-sentez yeteneği ve üst düzey bilişsel becerileri ölçmeye daha uygun olmalarına rağmen daha düşük güvenirlik, daha fazla zaman ve emek harcanması gibi olumsuz etkileri de beraberinde getirmektedirler. Bu çalışmada Eskişehir'deki bir dil okulunda eğitim gören 116 öğrencinin katılımıyla yabancı dil başarı ölçümünün açık uçlu ve çoktan seçmeli sorular kullanılarak karşılaştırılması amaçlanmıştır. Çalışmada dilbilgisi ve okuma becerileri dersi için hazırlanan iki açık uçlu sınavın çoktan seçmeli sorular içeren versiyonları hazırlanmış ve sonuçların karşılaştırılmasında kullanılmıştır. Araştırma sonuçları açık-uçlu ve çoktan seçmeli testler arasında madde zorluğu ve madde ayrımcılık düzeyleri açısından anlamlı bir fark olduğunu göstermiştir. Hem dilbilgisi hem de okuma çoktan seçmeli testlerinin açık uçlu testlerden daha kolay olduğu bulunmuştur.

To Cite This Article: Polat, M. (2020). Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Novitas-ROYAL (Research on Youth and Language)*, *14*(2), 76-96.

1. Introduction

Assessment in language teaching is a critical multidimensional task since measuring complex receptive and productive skills stand out as a natural and undeniable factor in controlling the operability of educational objectives. Normally, any formal assessment consists of at least two stages: the first step requires assessment, and the second step is the evaluation of existing measurements to make educational decisions. Beller and Gafni (1996) defined assessment as the observation of an attribute and representation of the result with numbers or other symbols, and they define evaluation as the process of handling the measurement results within a criterion and making a value judgement about the measured quality. Thus, assessment and evaluation form a kind of feedback mechanism regarding the efficiency and quality of the language teaching process. The outputs of this assessment and evaluation process provide important data about the quality of education, student performance and their current language proficiency levels; thus, it is a critical and dynamic process that enables the evaluation, correction and development of language education and its related dimensions. That is why assessment affects learning practices and defines educational objectives and curriculum in many ways (Cheryl et al., 2017).

A test is the first choice for instructors seeking to measure the students' achievement levels in a specific course where the desired cognitive ability of learners is evaluated through a number of test items (Brualdi, 1998); thus, a language test is a classical method of assessment. The content and aim of the test items determine whether the test succeeds in measuring the students' language performance or not. The attempts to test pre-planned educational outcomes using more valid and reliable methods has led to the development of various test techniques in foreign language assessment. While these assessment tools in education differ according to cognitive and affective factors, they are also categorized according to their test preparation techniques such as multiple choice, open ended, matching, true/false, and completion items (Black et al., 2003). Obviously, all these question types vary in their strengths and weaknesses, and none could be considered ideal for all language testing objectives (Cheryl et al., 2017). Nonetheless, considering all of these question types and their world-wide preference, popularity and extensive use in most language testing procedures, multiple choice and open-ended question types were studied in this research. Referring to the context of this study, multiple-choice question tests (MCQ) are frequently preferred to open-ended question tests (OEQ) in determining students' foreign language success including the high-stakes tests administered by the government in Turkey. Güler (2017) defines MCTs as tests where answers are not given by the respondents, the right answer is given by the test designer among the options, and the respondents are expected to find the correct answer. Such tests are widely used in language assessment as they are cheap, fast and easy to administer. However, they are also criticized since they are not adequate to measure many cognitive skills including critical thinking, creativity, motivation, flexibility, curiosity, perseverance, reliability, resilience, empathy, enthusiasm, discipline, self-awareness or self-direction (Brown, 2004).

On the other hand, OEQs require a full answer, based on the student's own knowledge or ideas. This type of question is generally subjective and does not direct the student to a certain answer. Supporters of using OEQ in language tests claim that open-ended items are advantageous in testing since they enable measuring higher-level foreign language skills (Badger & Thomas, 1992; Cooney et al., 2004). Akay et al., (2006) believed that by utilizing open-ended items, students can be asked to solve a real-life problem with some missing knowledge, which does not have a single (and fixed) solution, so that students can use reasoning and make some contributions through making assumptions and comments about the missing information.

Which question type in what kind of task and course all depend on the test maker, and this calibration requires expertise and experience in testing. Moreover, in foreign language testing, there are two important aspects for consideration: which cognitive levels are to be tested and what question types are to be used. First, a proper language test is classified as one which covers various cognitive levels to identify different abilities of language learners (Seddon, 1978); therefore, testing multiple cognitive levels with different questions enriches the validity of the language test. Next, the intended cognitive level of the test item to be measured is critical in choosing the question types; thus, determining the strengths and weaknesses of MCQs and OEQs could be useful to provide deeper insights to assess various cognitive skills. Thus, this study aims to investigate if there would be a difference in students' performance when the same questions are asked in MCQ and OEQ formats in different cognitive domain levels. Results of this study will be significant for language teachers and testing units since possible differences in lower cognitive domain levels between in MCQ and OEQ formats would guide test developers in their item-type selections for language tests.

2. Theoretical Framework

2.1. Multiple-choice versus Open-ended

Item format and its impacts on assessment quality has been an issue of discussion since the application of standardized language tests. Almost all international language tests such as TOEFL and IELTS use MCQs in their separate sections to test grammar, reading and listening in the most objective way. Klufa (2015) stated that MCQs have a number of advantages in testing; for instance, providing scoring reliability in crowded groups is easier with these tests and the ability to accommodate a large number of items allows it to cover critical content in the subject area along with high content validity. Especially in mass-testing where thousands (sometimes tens or hundreds of thousands) of students take the same test at the same time, MCQs are considered the most reasonable, reliable and affordable type of questioning (Rauch & Hartig, 2010). What is more, Turgut and Baykul (2012) reported that since MCQs do not include partial scoring, they provide more objective scoring options. As the scoring range shrinks, the reliability among the scores or test results increases (Case & Swanson, 2002). In addition to this, MCQs can give schools, teachers and students diagnostic results because they can be scored quickly (Bennett et al., 1991). With the help of optic readers, scoring a thousand students' papers takes just an hour or less along with the analysis showing the most confusing items or the easiest ones which were answered correctly by most students. These merits of MCQs help test designers not only to check the education program but also to edit the problematic items which were found too easy or too difficult for the students. This allows the test designers to rewrite or change those items or their distracters completely and convert the test into a better one.

Some researchers agreed that most MCQs do not involve higher thinking skills and include a potential of guessing (sometimes just by chance) the correct option which reduces the validity and reliability of the overall test (Breland et al., 1994; Crombach, 1988; Freahat & Smadi, 2014). Harrison et al., (2017) add to this stating that nowadays students have enough expertise to find the correct option even if they do not have enough knowledge about the task. In a similar vein, Cahill and Leonard (1999) found out that when learners are over tested by MCQs on tests, skilled testees perform well just because of their ability to recognize distracters. Martinez (1999) also underlined the same problem and stated that MCQs are weak at eliciting upper level cognitive processes which might misdirect teachers about the learners' success or mastery of the language content. Last but not least, Chan and Kennedy (2002) stated that in addition to decreasing the reliability because of tester guessing, MCQs may not spot higher levels of reasoning which also decreases the reliability of the test results.

Findings from the literature reveal that, like MCQs, an OEQ also has a number of pros and cons. According to Magliano et al., (2007), the most basic advantage of an OEQ is that it provides detailed responses from the learners which could increase the validity of the test. Particularly in the assessment of productive skills such as speaking or writing, using an OEQ is a wiser decision since higher order thinking and reasoning skills are required. Moreover, an OEQ eliminates the possibility of guessing the correct answer since limited information and no answer choices are provided with the questions. Lee et al., (2011) supported the same fact and reported that the chance factor in a MCQ and the possibility of finding the right answer by eliminating it is not the case for an OEQ. Talking about this advantage, it should also be mentioned that apart from their superiority in measuring high-level skills, OEQs are also advantageous since the answers obtained from open-ended items are more useful in determining and diagnosing the teaching process and its backwash effect in detail (Cooney et al., 2004).

However, apart from the possible difficulties in administration and doubts about the reliability of their results, an OEQ has a number of drawbacks besides its advantages in testing. The issue of reliability in scoring comes first. Cook and Myers (2004) warned that while grading OEQ, graders engage in various extensive semantic processes which cause bias in scoring, and unfortunately this is inevitable unless computers are trained to grade the papers. The human factor brings a lot of wash-back effects in grading OEQs, including more time and effort in grading, the necessity of grader justification, openness to simple errors related to hand-writing and a lower number of questions than MCQs which might be a threat to the content validity (Bastin & Van der Linden, 2003). Another important concern in comparing OEQs with the other question forms is the lack of "memory retrieval" while answering the question. Epstein (2007) defined memory retrieval as a cognitive process for students which becomes active when they see the questions in a language test. It is believed that while answering OEQs, this retrieval process is less active since there are limited memory igniters in the question root. However, as Ruit and Carr (2011) stated, in MCQ, this memory retrieval was observed to be higher than other question forms since both the question and the distractors (even if they were wrong) ring different bells in learners' minds and help them recall previously heard, read or studied information faster. All in all, findings from the literature reveal that both MCQ and OEQ have certain advantages and disadvantages in testing which should be weighed and considered accordingly by the test-makers. Moreover, there are a number of other variables including the cognitive domain levels and their operability with different test items to be considered to make a thorough comparison of using MCQ and OEQ in language tests.

2.2. Cognitive Domain levels in Testing

It is traditional in item-response theories of testing to search for the cognitive skills and their interaction with each other which might contribute to a learner's decision-making in a testing environment. These skills may sometimes be related to a number of linguistic features or a test-taker's personal and intellectual qualities which could result in different levels of understanding and response. Quite frequently, cognitive skills to be considered in language testing are related to Benjamin Bloom's Taxonomy (Eber & Parker, 2007). Bloom's taxonomy contains six different levels of cognitive skills that occur respectively while learning. In the original version, which was designed in 1956, the levels were: knowledge, comprehension, application, analysis, synthesis, and evaluation which were later reconsidered and renamed as remember, understand, apply, analyze, evaluate, and create (Reckase & McKinley, 1991).

Similar to the assessment objectives in other skills, the core aim in foreign language assessment is to develop functional and reliable assessment tools to test learners' language skills according to various stages of the aforementioned taxonomy. In a similar vein, Seddon (1978) defined a good language test as one which covers different cognitive levels to identify various skills of language learners. In classical and mono-dimensional tests, a test item simply refers to the first or a single step of the Bloom's taxonomy, though the taxonomy is made up of six inter-connected steps: the first three for low order and the next three for high order cognitive skills (Haris & Omar, 2015). Orey (2010) categorized the first three levels of Bloom's taxonomy as the "lower order thinking skills" which include remembering, understanding, and applying, while the others refer to the "higher-order thinking skills" which include analyzing, evaluating, and creating. Since the scope of this study is to compare MCQs and OEQs in different cognitive domain levels, lower order thinking skills were targeted to analyze with MCQs and OEQs. Eber and Parker (2007) defined the "remember" level as the procedure of retrieving specific information from memory coded as rote memory which involves testing simple facts, knowledge of major ideas, and memorizing. As the new information is integrated with the existing cognitive framework the level of "understand" occurs. A number of cognitive procedures such as interpreting, classifying, summarizing, and comparing could be observed in this stage (Paul et al., 2014). Finally, in the "apply" level there are two separate cognitive processes (Reckase & McKinley, 1991). First, executing occurs when the content of the question is familiar to the test-taker before the implementation occurs to solve the problem. Of these three levels, the level of "applying" is obviously more complex than others since it involves the ability to break the information in the test item into its parts and to analyze how the pieces are inter-connected with each other.

In a number of studies, comparing the use of MCQ sand OEQs in testing, it was found that the focus and the research objectives varied a lot. Namely Vasan et al., (2017) compared students' scores on MCQs and OEQs; Duran and Tufan (2017) compared students' views regarding MCQs and OEQs, and Ko (2010) compared both question forms using different genres and measured the differences according to the test content. However, there is little evidence on how students' answers differ when they answer the same question in MCQ and OEQ formats in different subjects and cognitive domain levels. Thus, the aim of this study is to compare language learners' test scores consisting of MCQs and OEQs based on the same items prepared for grammar and reading courses in terms their psychometric properties and different cognitive domain levels. With this aim in mind, the research questions of the study were as follows:

1. Is there a significant difference between the MCQs and OEQs in terms of the average item difficulty index (p)?
2. Is there a significant difference between the MCQs and OEQs in terms of the average item discrimination index (r)?
3. Is there a significant difference between the reliability coefficients of the scores obtained from the MCQs and OEQs?

4. Is there a significant difference between the mean scores of tests which have MCQs and OEQs?
5. Is there a significant difference between students' mean scores from the MCQs and OEQs in terms of different cognitive domain levels?

3. Methodology

3.1. Research Design and Participants

This quantitative study had an experimental research design and was carried out in a state university's language school between 2018 and 2019 in Eskişehir, Turkey. The experimental design was preferred since in this type of research one or more independent variables in a context are manipulated by the researcher and applied to one or more dependent variables to measure the possible effect(s) (Büyüköztürk, 2013). The effect of the independent variables on the dependent variables is usually tested and evaluated clearly since most of the possible variations were considered; therefore, researchers are able to draw reasonable conclusions in the end regarding the relationship between the variables. A total of 116 students (55 females, 61 males) from five different groups at the lower-intermediate level participated in this study voluntarily. They were not grouped as experimental or control groups since they all took the same tests throughout the study. All the participants were the students of a state university's language school. The age range among participants was 18 - 23, and their main faculties were mostly engineering, business administration and communication.

The course contents, course materials and syllabus were the same for all the participants since they were all B level students. There were four levels at the school starting from D level which stands for beginners, C for elementary, B for lower-intermediate and A for intermediate level. These levels were determined by the school administration and they do not correspond to CEFR levels since A level is defined as the lowest language level in CEFR, whereas it stands for the highest level in this language school. To score the students' tests, six English instructors who had at least ten years of experience working at the language school participated in the study, of which four were females, and two were males. In this language school, B level students take 20 hours of English classes a week, and an integrated approach is preferred in the language program rather than a skill-based language teaching approach.

3.2. Instruments

To collect data, first, two separate achievement tests consisting of OEQs were developed within the scope of the B level course book's grammar and reading parts. In the next phase, two other tests including MCQs were developed from the items included in the OEQ tests. In total, the participants took four tests, and they were asked to answer questions on the same subject matter in the form of OEQs and MCQs. The grammar tests' contents were related to the use of connectors (when-while) in simple present and simple past tenses, while the reading tests' content included detecting the main idea of a reading text and answering comprehension questions related to the text. The reading achievement tests which were used in the study had the same reading texts and the same question contents (e.g., the same main idea from a paragraph but in different test formats).

In calculating the content validity of the grammar and reading tests consisting of OEQs and MCQs, expert opinion was given. In total, nine experts (i.e., a professor, an associate professor, two assistant professors and five lecturers, all nine of whom held PhDs) in ELT commented on the content of the grammar and reading tests. There were 20 items in the original grammar test, and after the expert check, five items were excluded from the test, and two items were revised. For the reading test, out of 18 items, three were excluded, and four items were revised. Finally, both tests were reshaped having 15 items each.

To calculate the content validity of each test, Lawshe's (1975) content validity ratio formula was used. This formula is a linear conversion of a relative level of consensus on how many experts within a group rate a particular question item as "essential," and content validity of each test was found with the following formula:

$$CVR = \frac{n_e - (N/2)}{N/2}$$

Figure 1. Lawshe's (1975) content validity ratio formulae

In the formula, *CVR* stands for the content validity ratio; n_e is the number of group members from the expert team who decided that an item was "essential," and *N* is the number of all the members in the expert group. Lawshe (1975) determined the minimum CVR critical value for nine experts as 0.85 at 0.05 significance level and was revised by Wilson et al., (2012) as 0.889, and this value was taken as a reference. Accordingly, the Content Validity Index (CVI) calculated on the averages for reading OEQ test was 0.912, and the CVI calculated on the averages for the grammar OEQ test was found to be 0.948. Since both CVI values were bigger than the critical value 0.889, which was the content validity ratio specified for nine experts (CVI ≥ CVR), it was determined that content validity of both tests was statistically significant.

In the evaluation phase of the tests, OEQs were scored as 0, 1 or 2 in partial scores. An analytical scoring key was prepared for both grammar and reading tests, and these sample rubrics were used for scoring student answers related to the OEQs. Each partial score was grouped in itself as 2 points for exactly correct answers, 1 point for partially correct answers and 0 for false or empty responses. Possible grading reactions to students' responses were expressed in the rubrics with various examples. To calculate the CVR of both analytic rubrics for the grammar and reading OEQs' grading, expert opinion was again taken, and the same group of nine experts commented on the usability, validity and intelligibility of the rubrics. The CVI calculated according to expert scorings for the reading OEQ test rubric was 0.903, and the CVI calculated on expert scorings for the grammar OEQ test rubric was calculated as 0.921. Since both CVI values were bigger than the critical value 0.889, which was the content validity ratio specified for nine experts (CVI ≥ CVR), it was determined that content validity of both tests' rubrics was statistically significant.

Next, to validate the reliability of both rubrics to be used in the evaluation of the grammar and reading tests', OEQs' rubrics' inter-rater reliability degrees were calculated. Among the OEQ tests' answer sheets, 20 papers (i.e.,10 papers for grammar, 10 papers for reading) were selected randomly and again randomly selected two raters from the rater group (six English instructors) were invited to grade the papers using the analytical scales developed for the study. Pearson Correlation and Krippendorff's Alpha coefficient were utilized to find the inter-rater reliability degrees for both rubrics. The analysis revealed that there was statistically significant positive correlation between the scores of the two raters in terms of the items and overall test (for the grammar rubric: $r = .929$, $p < .01$, for the reading rubric: $r = .903$, $p <$

.01). The Alpha coefficient of Krippendorff for both rubrics was over 0.9, thus indicative of a very high level of agreement among raters (Krippendorff, 2011) which also indicated a high inter-rater reliability in scoring OEQs.

Finally, OEQs of both reading and grammar tests were converted to MCQs, and for this process, four experts in ELT (an associate professor, an assistant professor and two lecturers, all of whom held PhDs) gave their opinions. After the input from the experts, two items in the grammar test and four items in the reading test were revised, and those items were later re-checked and approved by the expert team. In terms of the MCQ, a four-option format was preferred for both grammar and reading tests since the related literature suggests four-option items for a moderate difficulty level (Alsawalmeh & Feldt,1994; Baştürk, 2014; Yaman, 2016). To illustrate the conversion, two sample questions from the grammar test are presented below:

*** Jason, can I have your charger when …………………?*
a. you will finish with it *b. you finished with it*
c. you finish with it *d. you've finished with it*

***You forgot your cell-phone charger at home and in the classroom, you discovered that a friend is charging his mobile phone, which is the same as yours. You'd like to borrow it for a while so you ask your friend kindly:*

…………………………………………………………………………………..……?

3.3. Data Collection

To complete the administrative process, the researcher attained the necessary official permissions from the language schools' administration. Two separate meetings in five classrooms were arranged for the testing sessions. The students were informed about the aim of the study, and that it was voluntary. Out of 134 students, 116 agreed to participate. After getting students' consent for the study, grammar and reading OEQ tests were firstly given to the study group. The sequence of testing was important to maintain the integrity of the tests; that is, students didn't find the answers on the MCQ options and use them in their own responses on the OEQs. In all tests 15 questions were asked. The grammar tests included 15 items to test the use of when/while, and the reading tests included three paragraphs with 15 items. A total of 75 minutes was given to students to complete both grammar and reading OEQ tests. After the participants completed the OEQ tests for both grammar and reading, a ten-minute break was given, and the students were asked not to talk about the test items in order not to color the other students' judgements. The proctors, who were also in the team of graders, didn't leave the classrooms during the breaks and managed some conversations with students on different topics to keep them busy during the break. In the following session, MCQ tests were given to the students, and the allocated time for the second session was again 75 minutes. In the evaluation phase, the grader group was given the exam papers. Each rater had about 40 papers (20 grammar and 20 reading papers), and they all agreed to complete the scoring in one week.

3.4. Data Analysis

After the papers were scored, the test results were computed by the researcher using multiple methods. Firstly, the Z test for content validity ratio comparisons was used to compare item difficulty and item discrimination indices since the Z-test is a hypothesis test to figure out if scores from an achievement test are valid or repeatable. Secondly, Cronbach Alpha and KR-20 values were used to compare reliability coefficients. Finally, the Feldt test, which can be measured with the formula proposed by Alsawalmeh and Feldt (1994) was used to test the difference between two Alpha coefficients. The formula to compare the reliability coefficients of OEQ and MCQ tests for reading and writing was as follows:

$$W = \frac{(1 - \alpha_1)}{(1 - \alpha_2)}$$

a1: Test 1 (OEQ) reliability coefficients
a2: Test 2 (MCQ) reliability coefficients

Figure 2. Alsawalmeh and Feldt's (1994) formula

It was suggested by Cronbach (1988), Alsawalmeh and Feldt (1994) that in cases where the (k - 1).(N - 1) value exceeds 1000, this formula can be used for dependent groups and the obtained W value should be compared with the F(N-1, N-1) degrees of freedom (k: number of items in the test). Therefore, it was decided to use this formula when the (k - 1).(N - 1) value was over 1000 [(15 - 1).(116 - 1) = 1610)] in the reliability comparison (for the OEQ and MCQ tests).

Before deciding on the tests to be used in the analysis of the data, the normality of the data, which is one of the parametric test assumptions, was tested. Since there were four separate tests and the tests were scored differently, all test scores were standardized by converting them into Z scores. Based on the obtained Z scores, extreme values were determined. Next, it was checked whether skewness and kurtosis values were in the range of +1 and 1, and the test results indicated that the scores did not show a significant deviation from the normal distribution. Finally, the analysis revealed that the null hypothesis of the Kolmogorov-Smirnov test was accepted in the distribution of OEQ and MCQ test scores of both courses (p > .05). There was a normal distribution in the data set; therefore, parametric tests, including t-test, two-way-ANOVA, and Pearson Correlation, were used in the analyses.

4. Findings

In this section, the findings obtained as a result of the analysis of the data through various statistical tests were included. The findings, namely grammar and reading tests, are stated under two separate headings. The findings are presented separately in the same order as the research problems were listed.

4.1. Grammar Test Results

Within the scope of the research, firstly, test statistics and item parameters (p, r) based on item scores of MCQ and OEQ tests were determined and the significance of the difference between them was tested.

Table 1

Comparison of grammar MCQ and OEQ test items' difficulty (p) and item discrimination (r)

Item	p1	p2	z	Cohen d	r1	r2	z	Cohen d
1	0.59^{u}	0.47^{u}	1.0	0.27	0.37^{u}	1^{u}	-9.90^{*}	1.87
2	0.31^{a}	0.95^{a}	-9.43^{*}	1.44	0.29^{a}	0.22^{a}	1.17	0.17
3	0.46^{a}	0.76^{a}	-4.36^{*}	0.65	0.41^{a}	0.91^{a}	-7.58^{*}	1.13
4	0.71^{u}	0.80^{u}	-1.82	0.29	0.58^{u}	0.68^{u}	-1.45	0.19
5	0.13^{r}	0.64^{r}	-7.74^{*}	1.15	0.34^{r}	1^{r}	-10.21^{*}	1.97
6	0.51^{a}	0.46^{a}	1.01	0.15	0.43^{a}	1^{a}	-9.25^{*}	1.77
7	0.65^{r}	0.76^{r}	-1.68	0.27	0.73^{r}	0.95^{r}	-3.85^{*}	0.51
8	0.77^{r}	0.86^{r}	-1.71	0.23	0.23^{r}	0.45^{r}	-3.21^{*}	0.45
9	0.34^{a}	0.37^{a}	-0.31	0.09	0.31^{a}	1^{a}	-10.1^{*}	1.94
10	0.12^{u}	0.70^{u}	-8.48^{*}	1.29	0.19^{u}	0.97^{u}	-11.39^{*}	1.82
11	0.40^{a}	0.66^{a}	-3.72^{*}	0.55	0.54^{a}	1^{a}	-7.93^{*}	1.53
12	0.68^{r}	0.51^{r}	2.49^{*}	0.37	0.23^{r}	0.94^{r}	-10.27^{*}	1.64
13	0.92^{u}	0.93^{u}	-0.23	0.06	0.22^{u}	0.33^{u}	-1.76	0.25
14	0.63^{u}	0.51^{u}	1.75	0.27	0.46^{u}	1^{u}	-8.66^{*}	1.66
15	0.52^{r}	0.77^{r}	-3.69^{*}	0.53	0.57^{r}	0.93^{r}	-5.97^{*}	0.83
Remembering	0.65	0.72	-1.86	0.15	0.44	0.81	-5.43^{*}	0.61
Understanding	0.46	0.63	-2.44^{*}	0.37	0.35	0.75	-5.72^{*}	0.82
Applying	0.51	0.62	-1.59	0.24	0.37	0.84	-6.86^{*}	0.93
Test mean	0.52	0.68	-2.18^{*}	0.31	0.39	0.83	-6.07^{*}	0.87

(*p < .05), Abbreviations, p: difficulty, r: item discrimination, ^{1}: OEQ test, ^{2}: MCQ test, r: Remembering, u: Understanding, a: Applying

When Table 1 is analyzed, item difficulty indices range between 0.12 - 0.92 for OEQ test and 0.37 - 0.95 for MCQ test in grammar assessment. Eber and Parker (2007) stated that the item difficulty index could be interpreted as .20 and below is very difficult; .21 - .40 is difficult; .41 - .60 is medium; .61 - .89 is easy; .90 and above is very easy. Therefore, in the OEQ test, there were two very difficult items, two difficult items, five medium items, five easy items and one very easy item; whereas, in the MCQ test, there were one difficult item, three medium items, nine easy items and two very easy items.

In terms of cognitive levels, it was found that OEQs and MCQs did not have a significant difference considering their item difficulties (p) in "remembering" and "applying" levels; however, in the "understanding" level, OEQs and MCQs had a significant difference. In this domain level, OEQs were harder than MCQs. Especially item ten, which was calculated as a very difficult question (p = .12) as an OEQ whereas the same question was an easy one as an MCQ (p = .70). Moreover, OEQs and MCQs' mean scores were significantly different from each other, and the grammar OEQ test was found to be more difficult than the grammar MCQ test. When the average difficulty level of the OEQ test was examined, it was seen that the test's difficulty was medium (p = .52); however, the MCQ test means showed that the test was easy (p = .68). Finally, it was determined that the difference between the difficulty levels of the tests was significant with a value of up to 0.31, and the test type variable had an average effect on the grammar test's difficulty.

In Table 1, the item discrimination indices range from 0.19 - 0.73 for the OEQ test and 0.22 - 1 for the MCQ test in grammar assessment. According to Wilson et al. (2012), the item discrimination index .19 and below is considered as very weak; between .20 - .29 is weak;

between .30 - .39 is acceptable; .40 - .59 is good, and 60 and above is very good. According to this scale, the OEQ test included one very weak item, four weak items, three acceptable items, six good items and one very good item. When the MCQ test was examined according to this item-discrimination scale; there were one weak, one acceptable, one good and twelve very good items. Considering the comparisons in term of item discrimination, it can be seen that, in all three cognitive domain levels, OEQ test items had lower item-discrimination levels than MCQ test items did. Item ten was again an outlier since its item discrimination was calculated as very weak (r = .19) as an OEQ whereas its item discrimination was very good (r = .97) as a MCQ.

When the item discrimination levels were analyzed in terms of overall tests, it can be said that the average discrimination level of the OEQ test was acceptable (r = .39), and the average discrimination level of the MCQ test was very good (r = .83). When the average discrimination levels of the tests were compared, it was found that the mean discrimination levels of the OEQ test and the MCQ test differed significantly, and the MCQ test questions were more distinctive than OEQ tests questions. Finally, it was determined that the difference between the discrimination levels of the tests was up to 0.86, and the test type variable had a great effect on test discrimination.

Table 2

Comparison of grammar OEQ and MCQ tests' reliability coefficients

Test type	**N**	**k**	**Cronbach Alpha/KR 20**	**W**	**Cohen d**
OEQ	116	15	.671	0.972	0.139
MCQ	116	15	.703		

When Table 2 is examined, the reliability coefficients for both tests were above 0.60. According to Ringim et al. (2012), the Cronbach Alpha value for a test may be as low as 0.60, and the test should still be considered a reliable tool for measurement. Other researchers also stated that the Cronbach Alpha value between .60 - .70 could be considered average and above .70 good (Hair et al., 2006). Accordingly, it can be interpreted that the reliability coefficients regarding both tests were at acceptable levels. In addition, according to Table 2, no significant difference was found between the reliability coefficients of the OEQ and MCQ tests [W <F (101,101) = 1.39]. Thus, it can be said that the test type variable is not an effective variable on reliability of the tests.

Table 3

Comparison of grammar OEQ and MCQ tests' scores

Tests	**N**	**Mean**	**SD**	**df**	**t**	**p**
OEQ	116	44.8	63.12	115	-2.961	.003*
MCQ	116	59.7	60.78			

(p < .05)

When Table 3 is analyzed, there is a significant difference between the scores that students received from the OEQ and the MCQ tests (t = -2.961, p < .05). It was found that students' grammar MCQ test's score averages were higher than their OEQ test's scores. This suggests students can be more successful in MCQ tests when compared to OEQ ones in grammar assessment.

Table 4

Comparison of grammar OEQ and MCQ test scores in different cognitive domain levels

Level	Tests	N	Mean	SD	df	t	p
Remembering	OEQ	116	48.9	23.9	115	-1.455	.149
	MCQ	116	52.4	22.1			
Understanding	OEQ	116	47.5	25.6	115	-1.339	.181
	MCQ	116	51.3	26.9			
Applying	OEQ	116	46.8	26.7	115	-1.682	.117
	MCQ	116	52.4	25.1			

(p< .05)

According to findings given in Table 4, there was no significant difference between students' OEQ and MCQ tests' scores in terms of cognitive domain levels: remembering, understanding and applying (t =. - 1.455, -1.339, -1.682; p > .05).

Table 5

Pearson Correlation test results of grammar OEQ and MCQ test scores in different levels

Levels	r	p	r2
Remembering *(OEQ * MCQ)*	.492**	.000	.23
Understanding *(OEQ * MCQ)*	.533**	.000	.31
Applying *(OEQ * MCQ)*	.418**	.000	.19

(p< .01)

Pearson Correlation in Table 5 showed that there was a moderate positive relationship between the scores of grammar OEQ and MCQ tests in remembering level (r = .492, p < .01, r2 = .23). Another moderate positive correlation was found between the scores in the understanding level (r = .533, p < .01, r2 = .31). In the applying level, it was again concluded that there was a moderate correlation among the scores (r = .418, p < .01, r2 = .19).

Table 6

Two-way ANOVA to compare grammar OEQ and MCQ test scores according to gender

	Variance	Sum of Squares	df	F	p
Remembering	Test type (*OEQ/MCQ)*	653.694	1	2.147	.149
	Gender	3081.921	1	3.954	.060
	Test type* Gender	39.668	1	.127	.741
Understanding	Test type (*OEQ/MCQ)*	2901.542	1	6.158	.013*
	Gender	251.765	1	.163	.701
	Test type* Gender	7.818	1	.017	.892
Applying	Test type (*OEQ/MCQ)*	775.963	1	2.043	.159
	Gender	958.948	1	.985	.327
	Test type* Gender	1817.216	1	4.776	.029*

(p< .05)

The ANOVA test results in Table 6 revealed that the main effect of the test type at the level of remembering [F (1,653.694) = 2.147, p > .05], the gender variable main effect [F (1,3081.921) = 3.954, p > .05] and the test type, gender effect [F (1,39.3668) =. 127, p > .05] were not significant. Thus, it can be concluded that OEQ and MCQ test scores in the remembering level of female and male students do not differ significantly. At the understanding level, the main effect of the test type is significant [F (1, 2901.542) = 6.158, p < .05]; however, gender main effect [F (1,251.765) =. 163, p > .05] and test type * gender effect [F (1,7.818) =. 017, p > .05] were not significant. Finally, when the level of applying is examined, the main effect of the test type [F (1,775.963) = 2.043, p > .05] and the gender main effect [F (1,958.948) =. 985, p > .05] were not significant but test type * gender effect was found to be significant [F (1,1817.216) = 4.776, p <.05]. Consequently, it was found that the female and male students' OEQ and MCQ test scores differed at the applying level.

4.2. Reading Test Results

In this part of the study, the results of the reading test's statistics and item parameters (p, r) based on item scores of MCQ and OEQ tests were shown.

Table 7

Comparison of reading MCQ and OEQ test items' difficulty (p) and item discrimination (r)

Item	p1	p2	z	Cohen d	r1	r2	z	Cohen d
1	0.59^u	0.91^u	-5.71*	0.89	0.44^u	0.24^u	3.24*	0.45
2	0.63^a	0.78^a	-2.12*	0.31	0.54^a	0.83^a	-4.40*	0.66
3	0.90^r	0.95^r	-1.46	0.15	0.15^r	0.13^r	0.5	0.07
4	0.76^r	0.95^r	-4.22*	0.63	0.52^r	0.12^r	6.37*	0.95
5	0.84^u	0.93^u	-1.96	0.23	0.32^u	0.29^u	0.33	0.05
6	0.60^a	0.91^a	-4.85*	0.65	0.72^a	0.36^a	5.19*	0.73
7	0.57^r	0.77^r	-3.05*	0.45	0.59^r	0.79^r	-3.08*	0.45
8	0.90^a	0.90^a	0	0	0.08^a	0.33^a	-4.52*	0.67
9	0.61^u	0.89^u	-4.46*	0.64	0.61^u	0.39^u	2.98*	0.43
10	0.66^a	0.95^a	-5.55	0.84	0.37^a	0.12^a	4.25*	0.63
11	0.82^r	0.90^r	-1.74	0.23	0.09^r	0.29^r	-4.34*	0.65
12	0.51^u	0.92^u	-6.44*	0.96	0.58^u	0.37^u	3.58*	0.53
13	0.58^u	0.84^u	-4.11*	0.63	0.28^u	0.56^u	-4.43*	0.65
14	0.77^a	0.56^a	3.34*	0.48	0.25^a	0.97^a	-10.8*	1.66
15	0.53^r	0.85^r	-4.96*	0.72	0.48^r	0.52^r	-0.58	0.11
Remembering	0.70	0.85	-2.61*	0.34	0.37	0.44	-1.04	0.13
Understanding	0.72	0.83	-2.04*	0.27	0.39	0.52	-1.88	0.27
Applying	0.69	0.81	-2.11*	0.56	0.35	0.62	-3.88*	0.58
Test mean	0.68	0.87	-2.23*	0.39	0.39	0.42	2.03*	0.33

(*p< .05), Abbreviations, p: difficulty, r: item discrimination, ^{1}: OEQ test, ^{2}: MCQ test, r: Remembering, u: Understanding, a: Applying

The results of the statistical analysis presented in Table7 show that item difficulty indices range from 0.51-0.90 for OEQ test and 0.56-0.95 for MCQ test in the reading assessment. Eber and Parker (2007) stated that the item difficulty index could be interpreted as .20 and below is very difficult; .21 - .40 is difficult; .41 - .60 is medium; .61 - .89 is easy; .90 and above is very easy. Therefore, in the OEQ test, there were six medium, seven easy and two very easy items, whereas; in the MCQ test there were one medium, five easy and nine very easy items.

In terms of cognitive levels, it was found that OEQs and MCQs had significant differences considering their item difficulties (p) in "remembering," "applying" and "understanding" levels. Moreover, OEQs' and MCQs' mean scores were significantly different from each other, and the OEQ reading test was found to be more difficult than the MCQ reading test. When the average difficulty level of the OEQ reading test was examined, it was seen that the test's difficulty level was easy (p = .68), and the MCQ reading test means showed the MCQ reading test was also an easy test (p = .87). Next, as the average difficulty levels of the tests were compared, it was found that the average difficulty levels of the OEQ test and the MCQ test differed significantly, and the MCQ reading test was an easier test compared to the OEQ reading test. Finally, it was determined that the difference between the difficulty levels of the tests was up to 0.39, and the test type variable had a small effect on the reading test's difficulty.

In Table 7, it could also be seen that item discrimination indices range between 0.08 - 0.72 for the OEQ test and 0.12 - 0.97 for the MCQ test in reading assessment. According to Wilson et al., (2012), the item discrimination index .19 and below is considered as very weak; between .20 - .29 is weak; between .30 - .39 is acceptable; .40 - .59 is good, and 60 and above is very good. With this in mind, in the OEQ test, there were three very weak items, two weak items, two acceptable items, six good items and one very good item in terms of item-discrimination. When the MCQ test was examined, there were three very weak, three weak, four acceptable, two good and three very good items in terms of item-discrimination.

Considering the comparisons in terms of item discrimination, in all three cognitive domain levels, OEQ test items had lower item discrimination levels than MCQ test items did. When the item discrimination levels were analyzed in terms of overall tests, the average discrimination level of the OEQ test was average (r = .39), and the average discrimination level of the MCQ test was good (r = .42). When the average discrimination levels of the tests were compared, the mean discrimination levels of the OEQ test and the MCQ test differed significantly, and the MCQ test questions were more distinctive than OEQ tests questions. Finally, it was determined that the difference between the discrimination levels of the tests was up to 0.33, and the test type variable had an average effect on test discrimination.

Table 8

Comparison of reading MCQ and OEQ tests' reliability coefficients

Test type	**N**	**k**	**Cronbach Alpha/KR 20**	**W**	**Cohen d**
MCQ	116	15	.683	0.964	0.138
OEQ	116	15	.674		

When Table 8 is examined, the reliability coefficients for both reading tests were above 0.60. Accordingly, the reliability coefficients regarding both tests were at acceptable levels. In addition, according to Table 2, no significant difference was found between the reliability coefficients of the OEQ and MCQ tests [W < F (101,101) = 1.38]. Thus, the test type variable is not an effective variable on reliability of reading tests.

Table 9

Comparison of reading OEQ and MCQ test scores

Tests	N	Mean	SD	df	t	p
OEQ	116	55.8	53.21	115	-3.213	.001
MCQ	116	72.7	49.52			

(p< .05)

When Table 9 is analyzed, there is a significant difference between the scores that students received from the OEQ test and the MCQ tests (t = -2.961, p < .05). Findings show that students' reading MCQ test's score averages were higher than their reading OEQ test's scores. Thus, we can conclude that students can be more successful in terms of academic scores in multiple-choice tests when compared to open-ended ones in the assessment of reading skills.

Table 10

Comparison of reading OEQ and MCQ test scores in different cognitive domain levels

Level	Tests	N	Mean	SD	df	t	p
Remembering	OEQ	116	50.3	22.3	115	-1.121	.243
	MCQ	116	53.1	20.9			
Understanding	OEQ	116	49.7	24.1	115	-1.439	.169
	MCQ	116	53.6	26.3			
Applying	OEQ	116	50.7	23.1	115	-1.463	.112
	MCQ	116	54.3	23.6			

(p < .05)

According to findings given in Table 10, there was no significant difference between students' OEQ and MCQ tests' scores in terms of cognitive domain levels: remembering, understanding and applying (t = - 1.121, -1.439, -1.463; p > .05).

Table 11

Pearson Correlation test results of grammar OEQ and MCQ test scores in different levels

Levels	r	p	r2
Remembering *(OEQ * MCQ)*	.521**	.000	.29
Understanding *(OEQ * MCQ)*	.547**	.000	.32
Applying *(OEQ * MCQ)*	.442**	.000	.20

(p< .01)

Pearson Correlation analysis in Table 11 showed that there was a moderate positive relationship between the scores of the reading OEQ and MCQ tests in remembering level (r = .521, p < .01, r2 = .29). Another moderate positive correlation was found between the scores in the understanding level (r = .547, p < .01, r2 = .32). In the applying level, it was again concluded that there was a moderate positive relationship among the scores (r = .442, p < .01, r2 = .20).

Table 12

Two-way ANOVA to compare grammar OEQ and MCQ test scores according to gender

	Variance	Sum of Squares	df	F	p
Remembering	Test type (*OEQ/MCQ)*	580.868	1	.621	.458
	Gender	21265.624	1	7.082	.011*
	Test type* Gender	1172.412	1	1.326	.278
Understanding	Test type (*OEQ/MCQ)*	212.540	1	.262	.633
	Gender	3686.825	1	2.635	.123
	Test type* Gender	147.664	1	.181	.688
Applying	Test type (*OEQ/MCQ)*	133.848	1	.190	.680
	Gender	11212.451	1	3.261	.088
	Test type* Gender	504.242	1	.702	.423

(p < .05)

The ANOVA test results in Table 12 revealed that the main effect of test type in the remembering level was not significant [(F (1,580.868) =. 621, p > .05]; the gender main effect was found to be significant [F (1,21265.624) = 7.082, p < .05], and the test type * gender interaction effect was not significant [F (1,1172.412) = 1.326, p > .05]. Accordingly, there is no significant difference between OEQ and MCQ tests' scores in the comprehension level of female and male students. At the understanding level, the main effect of the test type [F (1,212.540) =. 262, p > .05], the gender main effect [F (1,3686.825) = 2.635, p > .05] and the test type * gender interaction effect [F (1,147.664) =. 181, p > .05] were not significant for OEQ and MCQ test scores. Thus, there was no significant difference between female and male students' OEQ and MCQ test scores at the understanding level. Finally, at the level of applying, the main effect of the test type [F (1,133.848) =. 190, p > .05], the main effect of the gender [F (1,11212.451) = 3.261, p > .05] and the test type * gender interaction effects were not significant [F (1,504,242) = .702, p > .05]. Therefore, there was no significant difference between open-ended and multiple-choice test scores in the analysis level of female and male students.

5. Discussion and Conclusion

This study, carried out in Eskişehir between 2018 and 2019 with the participation of 116 voluntary university students studying at the language school of a state university, aimed to compare language learners' test scores on MCQs and OEQs based on the same items prepared for grammar and reading courses in terms of their psychometric properties and different cognitive domain levels. Research results showed that there was a significant difference between OEQ and MCQ tests in terms of item difficulty levels. Thus, it was observed that considering the mean score comparisons, MCQ tests for grammar and reading lessons were easier than OEQ tests and the effect of the test type was small considering the Cohen d values. Regarding the cognitive domain levels, there was a significant difference between grammar OEQ and MCQ tests. The difference in the understanding level might stem from the fact that the items in this level required a number of different cognitive skills and MCQ might have directed the students find the correct answers by using the multiple-choice format, by guessing the correct option or just by chance. In addition, it was determined that the items in the application level were easier for the students in the MCQ test. This might be due to the limitation of MCQ tests that allows the students to find the correct answer by guessing based on the given options, even though she/he did not know the correct answer exactly. It stands to reason that this characteristic of MCQs is a contributing factor to the findings supported in the literature that some students avoid writing or speaking in a foreign language unless they have to due to their negative attitudes towards productive skills and the fear of making simple mistakes (Black et al., 2003).

When the tests of reading were examined in terms of cognitive domain levels, it was determined that there was a significant difference in remembering, understanding and applying levels according to the test type. It was concluded that MCQ test items were easier than OEQ test items in all three cognitive domain levels. This might be due to the fact that OEQs require more complex cognitive skills as remembering, sequencing and expressing are necessary for answering. Namely, Koretz et al., (1993) stated that students mostly perceive OEQs as difficult ones and prefer to skip them without answering if they are given this option. Also, the study conducted by Duran and Tufan (2017) supports this interpretation. Their study revealed that students prefer having MCQ tests compared to OEQ tests since they spend less effort and time.

Research results show that there was a significant difference between OEQ tests and MCQ tests in terms of item discrimination degrees. Accordingly, it was found that both grammar and reading MCQ tests were more distinctive compared OEQ tests, and the type of test variable in grammar had a great effect on the item discrimination level whereas this effect was smaller in reading tests. When this difference was examined in terms of cognitive domain levels, it was determined that there was a significant difference between the item discrimination levels of grammar OEQ and MCQ tests in remembering, understanding and applying levels. Thus, it was concluded that the MCQ test was more distinctive than the OEQ test in all three cognitive domain levels. In their study, Taib and Yusoff (2014) reported the same finding and stated that the average item discrimination level of the language test including MCQs was higher than the test including OEQs.

Next, the results of the research revealed that there was no statistically significant difference between the reliability coefficients of both grammar and reading OEQ and MCQ tests. Therefore, this finding can be interpreted as the test type variable could not be an important variable that can make a significant difference on test reliability measures. Literature reveals that a number of variables such as being equipped with skills for quick answering, ability in narration, legible hand writing or a neat paper cause scoring bias which might reduce the reliability of scoring in OEQ tests (Bektaş & Kudubeş, 2014). What is more, the results of the research showed that there was a significant difference between students' OEQ and MCQ test scores, and the MCQ test results were higher so it was determined that the test type variable had a small effect on students' score difference. This difference might stem from the possibility of MCQs' options, chance factor, test habits and readiness of the participants for MCQ tests. This theory was also supported by other researchers as the options of MCQs might indirectly give students hints, and this might cause taking higher scores (Schuwirth et al., 1996). Also, Braun et al., (1990) stated that students can reach the correct answers in an MCQ test with the elimination method, but this is not the case with OEQs.

The results of the research showed that there was no significant difference between MCQ and OEQ test scores in remembering, understanding and applying levels in grammar and reading tests. Thus, this finding might imply that the performance of students in different cognitive domain levels do not change significantly according to the item type. In addition, the results of the research showed that there was a moderate and positive relationship among remembering, understanding and applying levels of OEQ and MCQ tests for both grammar and reading tests. Hence, it can be said that as the scores of the students in different cognitive domain levels of the OEQ test increase, their scores in the cognitive domain levels of the MCQ test would increase. Alternately, it is possible to state that as the scores in the cognitive domain levels of the OEQ test decrease, the scores in the cognitive domain levels of the MCQ test would decrease too. This finding reveals that the measurements in different cognitive domain levels do not differ in terms of test type, and a similar level of measurement could be gained. This finding was parallel with Hancock's (1994) findings. In his study, he found that there was a high level of correlation between the MCQ and OEQ tests in remembering, understanding and applying levels, and tests in different formats performed almost similarly in different cognitive domain levels.

Moreover, the results showed that there was no significant difference in terms of gender difference between the grammar and reading tests' scores regardless of the test format. Also, in terms of cognitive domain levels, it was determined that there was no significant difference in terms of gender difference between grammar and reading tests' scores in remembering, understanding and applying levels. Similarly, in the study conducted by Wright et al., (2016), no significant difference was found among the low-level cognitive level scores of different test types. Furthermore, research results showed that there was a significant difference between the tests in different formats developed to test the same objectives in terms of item difficulty and item discrimination. It was observed that students' grammar and reading MCQ test scores were higher than their OEQ test scores. It was also found that there was no significant difference between students' scores in the cognitive domain levels tested via OEQ and MCQ tests.

There are a number of limitations in this study. First, deeper and more reliable results could have been taken if more students had participated in this research. Next, participants were aware of the fact that this was an exploratory study, and the scores they would take from the grammar and reading tests would have no use in terms of assessment. Thus, it was not a real testing procedure and knowing this might have affected the participants' performances and the results of the comparisons. The last but not the least, just two sets of tests on reading and grammar were used. Making this comparison on more language skills including listening and vocabulary use could have given better insights, so using only grammar and reading tests is another limitation of this research.

To conclude, a number of research suggestions could be useful for future research on the same topic. To start with, the results of this research were based on a number of statistical comparisons between OEQ and MCQ tests in terms of their psychometric properties and students' performance, so further studies can be done to compare different test formats such as true-false, matching, fill-in the blanks etc. with multiple-choice or open-ended item types. In addition, low-level cognitive domain levels were highlighted in this study; thus, it would be a good idea to study high-level cognitive domain levels with different question types in foreign language assessment. Eventually, a mixed type research including statistical findings supported by students' opinions and feelings that could provide more data would be very useful to gain deeper insights about the research issue.

Ethical Issues

The author(s) confirm(s) that the study does not need ethics committee approval according to the research integrity rules in their country.

References

Akay, H., Soybaş, D., & Argün, Z. (2006). Problem kurma deneyimleri ve matematik öğretiminde açık-uçlu soruların kullanımı. *Kastamonu Eğitim Dergisi, 14(1)*, 129–146.

Alsawalmeh, Y. M., & Feldt, L.S. (1994). A modification of Feldt's test of two dependent alpha coefficients. *Psychometrika, 59*, 49–57. Retrieved in April, 2019 from: https://doi.org/10.1007/BF02294264

Badger, E., & Thomas, B. (1992). Open-ended questions in reading. *Practical Assessment, Research & Evaluation, 3*(4), 1991-1993.

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory. *Neuropsychology, 17*, 14–24. Retrieved in March, 2019 from: https://doi:10.1037/0894-4105.17.1.14

Baştürk, S. (2014). *Çoktan seçmeli testler. Eğitimde ölçme ve değerlendirme.* Nobel Akademik Yayıncılık.

Bektaş, M., & Kudubeş, A. A. (2014). Bir ölçme ve değerlendirme aracı olarak yazılı sınavlar. *Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi, 7*(4), 330-336.

Beller, M., & Gafni, N. (1996). International assessment of educational progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88,* 365-377.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice.* UK: McGraw-Hill Education.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27*(2), 93–108.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of advanced placement history examination. *Journal of Educational Measurement*, *31*, 275-293.

Brown, H. D. (2004). *Language assessment, principles and classroom practices.* USA: Longman.

Brualdi, A. C. (1998). Classroom questions. *Practical Assessment Research & Evaluation, 6*(6), 1-3. Doi: https://doi.org/10.7275/05rc-jd18

Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem.

Cahill, D. R., & Leonard, R. J. (1999). Missteps and masquerade in American medical academy: Clinical anatomists call for action. *Clinical Anatomy. 12*: 220-222

Case, S. M., & Swanson, D. B. (2002). Constructing Written Test Questions for the Basic and Clinical Sciences. *3rd Ed. Philadelphia, PA: National Board of Medical Examiners.* https://www.researchgate.net/publication/242759434_Constructing_Written_Test_Questions_For_the_Basic_and_Clinical_Sciences/link/00463529cfae562759000000/download

Cheryl, A. M., David, O. D., Bart, K., & Nagasawami, S. V. (2017). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *American Association of Anatomists Sci. Education, 11,* 254-261.

Cook, A. E., & Myers, J. L. (2004). Processing discourse roles in scripted narratives: Influence of context and world knowledge. *Journal of Memory and Language, 50,* 268-288. Doi: https://doi:10.1016/j.jml.2003.11.003

Cooney, T. J., Sanchez, W. B., Leatham, K., & Newborn, D. S. (2004). *Open-ended assessment in math: A searchable collection of 450+ questions.* "Open-ended Assessment in Math." www.heinemann.com/math

Cronbach, L. J. (1988). *Five perspectives on validity argument and test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Duran, E., & Tufan, B. S. (2017). The Effect of open-ended questions and multiple-choice questions on comprehension. *International Journal of Languages' Education and Teaching, 5(1)*, 242-254.

Epstein, R. M. (2007). Assessment in education. *New England Journal of Medicine. 3,* 387-396.

Freahat, N. M., & Smadi, O. M. (2014). Lower-order and higher-order reading questions in secondary and university level EFL textbooks in Jordan. *Theory and Practice in Language Studies, 4*(9), 1804-1813.

Güler, N. (2017). *Eğitimde ölçme ve değerlendirme* (10th Ed.). Ankara: Pegem Akademi.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th Ed.). Upper saddle River: Pearson Prentice Hall.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response tests. *The Journal of Experimental Education, 62*(2), 143–157.

Haris, S. S., & Omar, N. (2015). Bloom's taxonomy question categorization using rules and n-gram approach. *Journal of Theoretical & Applied Information Technology, 76*(3), 401-407.

Harrison, C. J., Konings, K. D., Schuwirth, L. W., Wass, V., & Van der Vleuten, C. P. (2017). Changing the culture of assessment: The dominance of the summative assessment paradigm. *BMC Medical Education. 17,* 73-87.

Klufa, J. (2015). Multiple choice question tests–advantages and disadvantages. *Mathematics and Computers in Sciences and Industry Journal*, *3*, 91-97.

Ko, M. H. (2010). A comparison of reading comprehension tests: Multiple-choice vs. open-ended. *English Teaching, 65*(1), 73-87.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from https://repository.upenn.edu/asc_papers/43

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). Omitted and not-reached items in mathematics in the 1990. *National Assessment of Educational Progress. 1*, 36-48.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575.

Lee, H-S., Liu, O., & 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. Doi: https://doi.org/10.1080/08957347.2011.554604

Magliano, J. P., Millis, K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies. Lawrence Erlbaum Associates Publishers, London (p. 107–136).

Martinez, M. E. (1999). Cognition and the question of test item format. *Education Psychology, 34,* 207-218.

Paul, D. V., Naik, S. B., & Pawar, J. D. (2014). An evolutionary approach for question selection from a question bank: A case study. *International Journal of ICT Research and Development in Africa (IJICTRDA), 4*(1), 61-75.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two dimensional IRT analysis. *Psychological Test and Assessment Modelling, 52*(4), 354–379.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Ringim, K. J., Razalli, M. R., & Hasnan, N. (2012). A framework of business process re-engineering factors and organizational performance of Nigerian banks. *Asian Social Sciences, 8*(4), 203.

Ruit, K., & Carr, P. (2011). Comparison of student performance on selected response versus constructed-response question formats in a medical neuroscience laboratory practical examination. *FASEB J, 2(15)*, 18-26.

Schuwirth, L. W., Van der Vleuten, C.P., & Donkers, H. (1996). A closer look at cueing effects in multiple-choice questions. *International Education Journal. 1(3)*, 44–49.

Seddon, G. M. (1978). The Properties of Bloom's Taxonomy of Educational Objectives for the Cognitive Domain. *Review of Educational Research, 48(2)*, 303-323.

Taib, F., & Yusoff, M. S. B. (2014). Difficulty index, discrimination index, sensitivity and specificity of long case and multiple-choice questions to predict students' examination performance. *Journal of Taibah University Medical Sciences, 9(2),* 110-114. Doi: https://doi.org/10.1016/j.jtumed.2013.12.002

Vasan, N. S., De Fouw, D. O., & Compton, S. (2011). Team based learning in anatomy: An efficient, effective and economical strategy. *Anatomic Science Education. 4*, 333-339.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45(3),* 197-210.

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE-Life Sciences Education, 15(2),* 1-23.

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi, 2(2),* 151-170.