Dil Eğitimi ve Araştırmaları Dergisi, 6(1), 216-237, April 2020
Journal of Language Education and Research, 6(1), 216-237, April 2020
https://dergipark.org.tr/tr/pub/jlere
Doi: 10.31464/jlere.584991

Araştırma Makalesi / Research Article

The Analysis of Multiple Choice Items in Final Exam of Preparatory Students

Hazırlık Öğrencilerine Uygulanan Final Sınavındaki Çoktan Seçmeli Maddelerin Analizi[*]

Sibel Toksöz** **Nazlı Baykal*****

Geliş / Received: 01.7.2019 **Kabul / Accepted:** 05.12.2019

ABSTRACT: The aim of this study is to examine the multiple choice final exams administered to the 210 non-compulsory preparatory school students. The study aims to analyze the exams in terms of three characteristics: item facility, item discrimination and distractor efficiency. The study had quantitative research design and the data were analyzed through Paired Samples T-Test and frequency analysis. The results of the study revealed that most items in final exams had moderate difficulty levels for the students. However, almost all items in the exams had low discrimination indices and some items had negative discrimination values. Furthermore, the results show that one third of the items in the exams had at least one non-functional distractor. At the end of the study, some guidelines were presented for teachers and test developers to make the items more functional.

Keywords: item analysis, item facility, item discrimination, distractor efficiency

ÖZ: Bu çalışmanın amacı 210 isteğe bağlı hazırlık öğrencisine uygulanan çoktan seçmeli final sınavlarını incelemektir. Çalışma sınavları çoktan seçmeli soruların üç niteliği yani madde kolaylığı, madde ayırıcılığı ve çeldirici yeteneği açısından incelemeyi amaçlamaktadır. Türk alanyazında çoktan seçmeli testlerle ilgili birçok çalışma bulunmasına rağmen çoktan seçmeli soruları madde analizi bakımından inceleyen çok az çalışma bulunmaktadır. Bu yüzden bu çalışma iki çoktan seçmeli sınavı madde analizi açısından inceleyerek alanyazına katkıda bulunacaktır. Bu çalışmada nicel araştırma yöntemi kullanılmıştır ve verileri analiz etmek için tek örneklem t-testi ve sıklık analizi kullanılmıştır. Sonuçlar göstermektedir ki, sınavlardaki soruların çoğu öğrenciler için orta zorluk derecesine sahiptir. Ancak, soruların hemen hepsi çok düşük madde ayırıcılığı değerine sahiptir ve bazı soruların ayırıcılık değerinin negatif olduğu bulunmuştur. Ayrıca, sonuçlar sınavlardaki soruların üçte birinde en az bir tane işlevsiz çeldirici olduğunu göstermektedir. Çalışmanın sonunda, soruları daha etkili hale getirmek için öğretmenler ve soru geliştirenler için bazı yönergeler sunulmuştur.

Anahtar sözcükler: madde analizi, madde kolaylığı, madde ayırıcılığı, çeldirici yeteneği

[*] This article is based on the author's MA thesis under the supervision of Nazlı BAYKAL at Süleyman Demirel University.

** Instructor, Burdur Mehmet Akif Ersoy University, School of Foreign Languages, sibelbayir@mehmetakif.edu.tr
***Associate Professor, Süleyman Demirel University, Faculty of Education, nazlibaykal@sdu.edu.tr

ISSN: 2149-5602

Introduction

Multiple choice tests are preferred by most of the teachers or institutions for a variety of disciplines in Turkey. At schools, the success in English has been determined mostly according to the multiple-choice (MC) exam results. Especially for higher grade levels and large scale testing programs MC tests are preferred for their ease and fastness in scoring (Rodgers & Harley, 1999). Since the tests have such an important role in determining the students' future academic careers or diploma grades the necessity for the tests being reliable, valid, efficient and functioning properly is becoming more crucial. "Since the quality of a test largely depends on the quality of the individual items" (Oluseyi & Olufemi, 2012, p.240), it seems significant to analyze the items before the test is given to the students. Item analysis is a general term and it is applied to investigate the test items for construction or revision (Oluseyi & Olufemi, 2012). With the help of item analysis, too easy or too difficult items can be identified and they can be dropped or in the same way, good items can be kept for future use. In the process of analyzing the test items, three types of indices can be calculated: Item facility (the difficulty level of the items), item discrimination (discriminatory power of the items between the high-achieving and the low-achieving students) and distractor efficiency (effectiveness of the distractors).

In spite of the extensive use of MC tests as mentioned above, up to now most of the studies in Turkey have focused on the usage (Temizkan & Sallabaş, 2011; Yaman, 2016), advantages or disadvantages of MC tests (Çalışkan & Kaşıkçı, 2010; Üstüner & Şengür, 2004; Yıldırım, 2010; Yaman, 2016). However, too little attention has been paid to MC tests in terms of item analysis in Turkish context. To name a few, Atalmış (2014) investigated whether item facility and item discrimination change when NOTA (None of the Above) option is added to the questions as a distractor and whether the test reliability changes when the number of the option decreases from 4 to 3. 1.130 students from sixteen differrent schools in Turkey and 100 students from one school in U.S.A participated to the study. Moreover, Toksöz and Ertunç (2017) analyzed a 50-item MC exam administered to 453 students studying at a state university in Turkey. Still, it seems prudent to examine the quality of the MC items, which is the primary aim of the present study. The study specifically aims to analyze the exam in terms of three characteristics of multiple-choice questions (MCQs): item facility, item discrimination and distractor efficiency. Bearing these aims in mind, this study attempts to respond to the following research questions:

1. . What is the difficulty level (item facility) of each item on the final exam test administered to non-compulsory preparatory school students?
2. What is the discrimination index (item discrimination) of each item on the final exam test administered to non-compulsory preparatory school students?
3. What is the distribution of the response patterns (distractor efficiency) for each of the options on the final exam test administered to non-compulsory preparatory school students?

This study is limited to 210 participants since the number is small, the generalizability of the findings to a larger number can be argued as one of the limitations of the study. Furthermore, the reader should bear in mind that this study is based on the three main characteristics of the MC test items: item facility, item discrimination and distractor efficiency. Therefore, it is beyond the scope of this study to present a full analysis of the items.

Item Analysis

An item "has always been the basic building block of a test" (Wainer, 1988, p. 2). Items are important in the sense that they play significant roles in improving the reliability of tests (Burton, 2004). As Coombe, Folse, and Hubley (2007) advocate to make a test function well, all the other essential parts such as items, keys and the distractors need to work effectively.

The main purpose of item analysis evaluating the test as a whole and analyzing the items individually is to construct and revise the test (Cechova, Sedlacik, Neubauer, 2014; Oluseyi & Olufemi, 2012). Coniam (2009) states that item analysis investigates how much each item contributes to the test's worth. Therefore, it could be inferred that item analysis provides much valuable and empirical data to the teachers or researchers about how the items in the test are performing (Oluseyi & Olufemi, 2012, p. 240). Useful implications and insights could be drawn from an item analysis for test developers and misleading or ambiguous items could be eliminated from the test or they might be improved for future use (Bodner, 1980; Oluseyi & Olufemi, 2012). Meanwhile by discarding the flawed items or revising them the quality of the test as a whole is improved (Hamzah & Abdullah, 2011; Olufemi & Oluseyi, 2012; Oppenheim, 2002).

Considering the literature on the subject, it is observed that, the studies on item quality are usually conducted on large-scaled standardized tests rather than classroom assessment (Stiggins & Bridgeford, 1985). Bodner (1980) pays attention to the lack of the studies in terms of item analysis stating that although multiple choice tests yield a lot of statistical data which are somewhat important and useful for the researchers they are mostly ignored. More empirical studies ought to be conducted on item analysis to improve tests and exams for future use and, thereby serving the testing aims.

Item Facility

The Difficulty index (DIF I) or item facility (IF) is symbolized as "*p". The *p*-value can range from 0.00 which means that nobody answered the item correctly, to 1.00 which means that everybody chose the correct option (DiBattista & Kurzawa, 2011; Oluseyi & Olufemi, 2012). When the value of DIF is big, it means it is an easy item; and if the item has a small value of DIF index that means the item is difficult (Gajjar, Rana, Kumar, Sharma, 2014; Oluseyi & Olufemi, 2012). With respect to the accepted difficulty ranges, there have been different cut-off points suggested by researchers such as .31 and .60 (Gajjar, et al., 2014); .30 and .92 (Jafarpur, 1999); .15 and .85 (Brown, H.D., 2004); .50 and .80 (Hamzah & Abdullah, 2011); .20 and .90 (Olufemi & Oluseyi, 2012); .30 and .80

(Coniam, 2009; Oppenheim, 2002); .30 and .70 (Brown, J.D., 2003); .50 and .90 (Haladyna & Downing, 1993).

In analyzing item facility the aim is not to find very difficult questions. If a test is too difficult might be unable to discriminate the students having different abilities (Coniam, 2009). According to Coombe et al. (2007) "ideal tests have a mix of difficulty levels…" (p. 163). According to Hamzah and Abdullah (2011) items with an average degree of difficulty can contribute to the reliability of a test. "Optimum test reliability demands more than just lengthy tests with non-overlapping questions; it also demands moderately difficult questions containing equally plausible distractors, plus (nevertheless) a high average score" (Bush, 2006, p. 400). Oppenheim (2002) advocates that there could be some items that are very easy if they are testing a well-known fact about the topic, however the number of those easy items should be limited.

Toksöz and Ertunç (2017) analyzed a 50-item multiple-choiced midterm exam administered to 453 students studying in language preparation classes and the results showed that 41 of the items had moderate difficulty levels ranging between .24 and .85. Moreover, 2 of the items were found to be very easy for the students having low difficulty indices (.11 and .07). Furthermore, they found that 7 of the items were too difficult having high difficulty values ranging between .86 and .98.

Item Discrimination

In item discrimination the students' performance on a test item is compared with their performance on the whole exam (Coombe, et al., 2007). Therefore, the focus in on U-L Index, "U" stands with the upper group of the test-takers and "L" stands for the lower group of the test-takers (Burton, 2001). To obtain item discrimination values, the most successful 30% of the answer papers and the least successful 30% of the papers are taken into consideration (Brown, 2004). Papers with intermediate scores are ignored.

To Brown, H. D. (2004), a highly discriminating item has a value close to perfect 1.0 and if an item fails to discriminate between the high-achieving and the low-achieving students, that means it has a value closer to zero; if the value is zero it means that this item couldn't discriminate at all. The maximum value 1 is obtained when a question is answered correctly by all of the high achieving students (the upper group) and by none of the low achieving students (the bottom group) and a negative value is obtained if the item is answered correctly mostly by the low achieving students (Burton, 2001; Hamzah &Abdullah, 2011).

An item may also have a negative discrimination value. That happens when high-achieving students cannot choose the correct option while low-achieving students can find the correct option. This may be because of that high-achieving students may interpret the question more difficult than it actually is and might be suspicious (Coombe et al., 2007; Gajjar, et al., 2014) or it might be just because of the complex wording or structure of the item (Gajjar, et al., 2014). In all cases, that item needs revision since those kinds of situations are undesirable for both teachers and students.

DiBattista and Kurzawa (2011) argue that the items with very low or very high discrimination values are likely to be problematic. Likewise, Reid (as cited in DiBattista &

Kurzawa, 2011) asserts that " even more problematic are items that function so poorly that they have a negative discrimination coefficient, perhaps because the wording is unclear or because two options rather than one are correct" (p.2). "Such items with negative DI are not only useless; but they actually serve to decrease the validity of the test" (Gajjar, et al., 2014, p. 19). DiBattista and Kurzawa (2011) state that the discrimination coefficient of a multiple choice exam must be a positive value, otherwise a MC item fails to function effectively.

In their study with 453 students Toksöz and Ertunç (2017) found that 14 items out of 50 had moderate item discrimination indices (.50 and higher). Moreover, they found that 36 of the items had low item discrimination values (.50 and lower). Also, one item was found to have a negative item discrimnaiton value (. -09). They claimed that this item had the potential to create a negative washback effect for the students.

Distractor Efficiency

Analysis of distractors separates the functional distractors which are chosen by some test takers and non-functional distractors which are seldom chosen by the test takers (Malau-Aduli & Zimitat, 2012). Distractors ought to look like correct answers for the students who did not understand the topics on the test (Coombe, et al., 2007). Moreover, distractors "reflect the points in an argument when a student's reasoning goes awry" (Buckles & Siegfried, 2006, p.52).

The frequencies showing the distribution of the responses can be benefitted to make a conclusion about the efficiency of a distractor. If a distractor is not chosen by most of the test takers even by the low achieving group that means that this distractor does not fool anyone. According to Downing and Haladyna (1997) "…at least 5% of examinees should select each of an item's distractors" (p.3). Similarly, Gajjar et al. (2014), and Ware and Vik (2009) define a distractor as non-functioning distractor (NFD) if the distractor is chosen by <5 % of the test takers. Nonfunctional distractors should be either replaced with a functioning one or be omitted from the test completely (Haladyna & Downing, 1989).

Tarrant, Mohammed, Ware (2009) observed that it was challenging enough to develop four functional distractors in five-option items. Moreover, there are a lot of distractors which are not functioning properly on classroom tests (DiBattista & Kurzawa, 2011). One way to write strong distractors might be to use fewer options; for instance, to use three-options instead of four-options (Haladyna, Downing, Rodriguez, 2002; Rogers & Harley, 1999; Bruno & Dirkzwager, 1995). Although four distractors have been regarded as a standard and common practice in MCQ tests (Bruno & Dirkzwager, 1995), and favored by teachers and examinees, researchers suggest that three functional distractors are more realistic and manageable besides being easier to prepare (Haladyna et al., 2002; Tarrant et al., 2009; Costin, 1970). Most studies (Ebel, 1969; Haladyna & Downing, 1993; Tversky, 1964; Bruno & Dirkzwager, 1995) have advocated three-option items instead of four highlighting that three-option items are as reliable as four or five alternatives. Three-option items can also provide some advantages to the teachers such as spending less time while forming the distractors (Tarrant et al., 2009) which may be argued to be one of the disadvantages of MC tests (Coombe et al., 2007).

Research claims that there is a crucial need to do item analysis of the multiple choice exams to enable more quality and functioning items for students and more accurate and reliable results for teachers or test developers (DiBattista & Kurzawa, 2011; Jafarpur, 1999; Goodrich, 1977; Rodger & Harley, 1999; Burton, 2001). However, there seems to be a gap in item analysis of multiple choice tests in Turkish literature. Hence, the study aims to analyze a multiple choice final exam administered to non-compulsory preparatory students at a state university.

Methodology

Participants and Data Collection Process

The study was conducted on 210 non-compulsory preparatory school students. The students were from different parts of Turkey and they were studying in different departments such as Engineering, International Trade, and Tourism and Hotel Management. In their weekly schedule in preparatory classes, the students were taking English lessons 20 hours a week: 10 hours of Main Course, 6 hours of Reading and Writing, and 4 hours of Grammar courses. The data were collected through the final exams administered to the non-compulsory preparatory school students. After the appropriate institutional permissions were secured, the quantitative data was collected right after the exam had been administered. To collect quantitative data, the final exam session I and the final exam session II were used.

Instruments (Final exams)

The final exam consists of four main parts: listening, grammar, vocabulary, reading (reading texts cloze test, conversation, situation, translation). The detailed information about the content of the exam is presented in Table 1 below.

Table 1. General Content of the Final Exams Administered to Preparatory Students

Part	Number of Items	Part	Number of Items
Listening	10	Reading	20
Grammar	20	Dialogue	4
Vocabulary	20	Situation	5
Cloze Test	10	Translation	6

The questions had equal points in the overall score. Listening questions had three-options; however, all the questions in the other parts had five-options. The students were not penalized for wrong answers. They got 1 point for each of their correct answers and 0 point for their incorrect answers. The questions in the other parts of the exams were directly used by the researcher without modification.

The exams were held in two different sessions according to their times. The first session was held at 12:45 and the second session was held at 15:15. The questions in each session were constructed with different but parallel questions. In other words, the same lexical and grammatical items were tried to be asked in both sessions. Also, the instructors proctored during the exam to prevent the students from cheating.

Data Analysis

While analyzing the quantitative data the students not selecting any of the options were eliminated to reach more accurate results. Therefore, 210 exam papers were taken into consideration for the statistical analysis although 266 students had taken the exam. To analyze the quantitative data, test takers' responses for each item on the final exams were analyzed through the statistics program IBM SPSS Version 20. During the data analysis the researcher focused on three main item characteristics: item facility, item discrimination and distractor efficiency. The quality criteria and the formulas for each of the quality indicators (item facility, item discrimination and distractor facility) were derived from Brown (2004).

Results

Item facility indices of the items on the final text exams

In this part, the item facility (IF) or difficulty (DIF) indices of final exam session 1 and final exam session 2 will be presented in tables and analyzed.

According to Table 2 below, three items (3 %) in final exam session 1 were too easy (p≥ 85) according to H.D. Brown's (2004) cut-off points.

Table 2. Item Facility (IF) indices of too easy items in final exam session 1

Item #	*p*	Item #	*p*	Item #	*p*
Item # 1	.85	Item # 6	.89	Item #8	.94

Table 3 below shows that eight items (8 %) in final exam session 1 were too difficult (<15) according to Brown's (2004) benchmark values. The rest of the items (57 %) in final exam session1 have moderate difficulty levels (between .15 and .85) according to Brown's (2004) cut-off points.

Table 3. Item Facility (IF) indices of too difficult items in final exam session 1

Item #	*p*	Item #	*p*	Item #	*p*
Item # 10	.10	Item # 13	.12	Item # 28	.12
Item # 32	.15	Item # 70	.09	Item # 72	.13
Item # 75	.13	Item # 94	.14		

According to Table 4 below, two items (2 %) in final exam session 2 are too easy (≥ 85) according to H.D. Brown's (2004) benchmark values.

Table 4. Item Facility (IF) indices of too easy items in final exam session 2

Item #	*p*	Item #	*p*
Item # 8	.92	Item # 83	.85

According to Table 5, 11 items (11.5 %) in final exam session 2 were too difficult (<15) in line with Brown's (2004) benchmark values. The rest of the items (86 %) in final exam session 2 had moderate difficulty levels (between .15 and .85) according to H.D. Brown's (2004) cut-off points.

Table 5. Item Facility (IF) indices of too difficult items in final exam session 2

Item #	*p*	Item #	*p*	Item #	*p*
Item # 6	.15	Item # 9	.03	Item # 11	.14
Item # 45	.14	Item # 49	.12	Item # 57	.05
Item # 62	.15	Item # 64	.10	Item # 71	.14
Item # 78	.10	Item # 89	.14		

Item discrimination indices of the items on the final test exams

In this part, the item discrimination indices (DI) of the items in the final exam session 1 and 2 will be presented in tables and analyzed.

Table 6 below shows that 6 items (6.31%) in final exam session 1 had negative item discrimination indices which means those items were answered correctly mostly by low achieving students (Burton, 2001; Hamzah & Abdullah, 2011). To illustrate, the distribution of responses for item # 60 in final exam session 1 is shown in Table 6 below.

Table 6. Items with negative discrimination indices in final exam session 1

Item #	*DI*	Item #	*DI*	Item #	*DI*
Item # 4	-0.01	Item # 6	-0.01	Item # 55	-0.02
Item # 60	-0.06	Item # 75	-0.02	Item # 94	-0.04

Table 7 below demonstrates that item # 60 gathered more correct answers from low ability students rather than high ability students.

Table 7. Distribution of responses for item # 60 in final exam session 1

Item # 60	#Correct	*#Incorrect*
High Ability Ss (Top 36)	10	26
Low Ability Ss (Bottom 36)	15	21

Table 8 below demonstrates that 6 items (6.31%) in final exam session 1 had moderate discrimination indices (Coombe, et al., 2007). However, Brown (2004) suggests that moderate level of discrimination should be .50 and above. In that respect none of the items could be argued to discriminate well at all. Moreover, final exam session 1 does not seem to meet the discrimination requirements suggested by Ware and Vik (2009) who suggest that greater or equal to 60% of the items should have moderate discrimination indexes. The remaining 83 items (87.36 %) had discrimination values which are zero or very low (<30) thereby these items were unable to meet the requirements.

Table 8. Items with moderate discrimination indices in final exam session 1

Item #	*DI*	Item #	*DI*	Item #	*DI*
Item # 15	.38	Item # 36	.38	Item # 48	.37
Item # 61	.33	Item # 91	.37	Item # 93	.30

According to Table 9 below, 7 items (7.36 %) had negative discrimination indexes which mean those items gathered more correct answers from high ability students rather than low ability students (Burton, 2001; Hamzah & Abdullah, 2011). To illustrate, the distribution of responses for item #6, #17, and #57 having negative discrimination indexes in final exam session 2 are shown below in Table 10, and Table 11 respectively.

Table 9. Items with negative discrimination indices in final exam session 2

Item #	*DI*	Item #	*DI*	Item #	*DI*
Item # 6	-0.07	Item # 9	-0.01	Item # 12	-0.02
Item # 17	-0.13	Item # 57	-0.04	Item # 60	-0.02
Item # 62	-0.04				

Table 10 below shows that most high ability students failed to answer item # 6 correctly in the final exam session 2. Low ability students were more successful for that item contrary to the expectations.

Table 10. Distribution of responses for item # 6 in final exam session 2

Item # 6	#Correct	*#Incorrect*
High Ability Ss (Top 34)	5	29
Low Ability Ss (Bottom 34)	10	14

As seen in Table 11 below, more students from low ability group rather than high ability group were able to answer item # 17 correctly in final exam session 2. These items having negative item discrimination indices are probale to create negative washback effect for high ability students (Hughes, 2003). Hence, they need to be revised or modified by test developers or teachers.

Table 11. Distribution of responses for item # 17 in final exam session 2

Item # 17	#Correct	*#Incorrect*
High Ability Ss (Top 34)	9	25
Low Ability Ss (Bottom 34)	18	16

Table 12 below demonstrates that 8 items (8.42 %) in final exam session 2 had acceptable discrimination indexes (.30 and above) (Coombe, et al., 2007). However, Brown, H.D. (2004) suggest that o moderate level of discrimination should be .50 and above. In that respect none of the items could be argued to discriminate well at all. Moreover, the final exam session 2 does not seem to meet the discrimination requirements suggested by Ware and Vik (2009) who suggest that greater or equal to 60% of the items should have moderate discrimination indexes.The remaining 80 items (84.21 %) had discrimination values which are zero or very low (<30) thereby these items were unable to meet the requirements.

Table 12. Items with moderate discrimination indices in final exam session 2

Item #	*DI*	Item #	*DI*	Item #	*DI*
Item # 43	.30	Item # 47	.35	Item # 72	.32
Item # 86	.32	Item # 92	.33	Item # 93	.33
Item # 94	.32	Item # 95	.30		

Distractor efficiency of the options of the items on the final test exams

A distractor is defined as non-functioning distractor (NFD) if the distractor is chosen by <5 % of the test taker (Gajjar et al. 2014; Ware & Vik, 2009; Downing & Haladyna, 1997). In this part, the distribution of the responses of the items having non-functional distractors (NFD) in final exam session 1 and final exam session 2 will be presented in tables and analyzed. Table 13 below shows that 3 items (30%) in the listening part of final exam session 1 had NFD distractors.

Table 13. Items with NFDs in the listening part of final exam session 1

Item #	A	B	C	*D*	E
# 1	11	5	**92**	-	-
# 6	3	**97**	8	-	-
# 8	4	2	**102**	-	-

Note. Bold options are the correct answers

Table 14 below demonstrates that 8 items (40%) in the grammar part of final exam session 1 had NFD distractors.

Table 14. Items with NFDs in the grammar part of final exam session 1

Item #	A	B	C	*D*	E
# 12	25	34	**30**	3	16
# 15	22	**43**	5	33	5
# 16	**41**	22	11	29	5
# 19	**74**	21	7	3	3
# 21	16	6	15	**70**	1
# 25	9	41	4	9	**45**
# 29	19	11	12	5	**61**
# 30	**48**	24	19	14	3

Note. Bold options are the correct answers

Table 15 below shows that 5 items (25%) in the vocabulary part of final exam session 1 had NFD distractors.

Table 15. Items with NFDs in the vocabulary part of final exam session 1

Item #	A	B	C	*D*	E
# 35	6	7	**87**	4	4
# 36	32	4	5	3	**64**
# 37	18	**72**	13	2	3
# 38	11	10	3	**73**	11
# 40	23	33	15	**32**	5

Note. Bold options are the correct answers

Table 16 below shows that 5 items (50%) in the cloze test part of final exam session 1 had NFD distractors.

Table 16. Items with NFDs in the cloze Test part of final exam session 1

Item #	A	B	C	*D*	E
# 51	16	**46**	15	26	5
# 54	**85**	3	7	6	7
# 55	18	4	30	**36**	20
# 56	**38**	38	17	10	5
# 57	24	**27**	23	32	2

Note. Bold options are the correct answers

Table 17 below demonstrates that 2 items (33.3%) in the translation part of final exam session 1 had NFD distractors.

Table 17. Items with NFDs in the translation part of final exam session 1

Item #	A	B	C	*D*	E
# 61	16	13	4	**62**	13
# 65	5	10	**61**	10	22

Note. Bold options are the correct answers

Table 18 below shows that 5 items (25%) in the reading part of final exam session 1 had NFD distractors.

Table 18. Items with NFDs in the reading part of final exam session 1

Item #	A	B	C	*D*	E
# 67	7	3	**62**	32	4
# 72	**15**	8	67	14	4
# 83	**65**	13	22	6	2
# 84	7	**83**	4	7	7
# 86	3	**56**	33	10	6

Note. Bold options are the correct answers

All the distractors in the dialogue and the situation parts of final exam session 1 were found to be functional. That means they all the items in these parts were chosen by more than 5% of the test takers who took finel exam session 1. Hence, the tables of these parts are not presented here. Overall, 10% of the distractors in the final exam session 1 were found to be flawed because they were chosen by less than 5% of the examinees who took final exam session 1. In all, 29% of the items had at least one of these flawed distractors in the final exam session 1. 90% of the distractors were found to function properly.

Table 19 below shows that 2 items (20%) in the listening part of final exam session 2 had NFD distractors.

Table 19. Items with NFDs in the listening part of final exam session 2

Item #	A	B	C	*D*	E
# 8	**94**	4	4	-	-
# 9	89	**4**	9	-	-

Note. Bold options are the correct answers

Table 20 below demonstrates that 4 items (20%) in the grammar part of final exam session 2 had NFD distractors.

Table 20. Items with NFDs in the grammar part of final exam session 2

Item #	A	B	C	*D*	E
# 17	6	41	**45**	2	8
# 18	19	20	11	**51**	1
# 25	**56**	12	10	3	21
# 26	**22**	15	47	1	17

Note. Bold options are the correct answers

Table 21 below shows that 3 items (15%) in the vocabulary part of the final exam session 2 had NFD distractors.

Table 21. Items with NFDs in the vocabulary part of final exam session 2

Item #	A	B	C	*D*	E
# 41	17	6	2	**70**	7
# 47	**56**	33	8	3	2
# 49	11	36	4	38	**13**

Note. Bold options are the correct answers

Table 22 below demonstrates that 2 items (20%) in the cloze test part of the final exam session 2 had NFD distractors.

Table 22. Items with NFDs in the cloze Test part of final exam session 2

Item #	A	B	C	*D*	E
# 51	**71**	12	4	4	11
# 53	**47**	4	7	32	12

Note. Bold options are the correct answers

Table 23 below shows that only 1 item (5%) in the reading part of the final exam session 2 had NFD distractors.

Table 23. Items with NFDs in the reading part of final exam session 2

Item #	A	B	C	*D*	E
# 65	**37**	8	32	21	4

Note. Bold options are the correct answers

Table 24 below demonstrates that 2 items (33.3%) in translation part of final exam session 2 had NFD distractors.

Table 24. Items with NFDs in the translation part of final exam session 2

Item #	A	B	C	*D*	E
# 82	38	**46**	3	4	11
# 83	4	3	**87**	5	3

Note. Bold options are the correct answers

Table 25 below shows that 2 items (22.2%) in the dialogue part of the final exam session 2 had NFD distractors.

Table 25. Items with NFDs in the dialogue part of final exam session 2

Item #	A	B	C	*D*	E
# 87	13	**76**	5	6	2
# 91	30	17	20	**33**	2

Note. Bold options are the correct answers

All the distractors in the situation part of final exam session 2 were found to be functional. That means the distractors in this part of the exam were chosen by more than 5% of the test takers who took final exam session 1. Hence, the table related to this part is not presented here. Overall, nearly 6 % of the distractors were found to be flawed because they were chosen by less than 5% of the examinees who took final exam session 2. In all, 16.8% of the items had at least one of these flawed distractors in the final exam session 2. Almost 93% of the distractors were found to function properly.

Conclusion and Discussion

Research Question 1: What is the difficulty level (item facility) of each item on the final exam test administered to non-compulsory preparatory school students?

The findings revealed that most of the items in final exams had moderate difficulty levels. These items seem to be ideal and appropriate for the students' levels and they need no modification; therefore, they could be maintained and used in future exams (Ebel, 1967). These items having moderate difficulty levels can also contribute to the reliability of the exam as a whole (Bush, 2006; Brown, J. D. 2003). These items are valid since they showed that the students learned the content measured by these items (Malau-Aduli & Zimitat, 2012). Furthermore, the items having moderate difficulty levels might be argued to serve the aims of testing. Therefore, test developers and teachers should be trying to write items with appropriate difficulty levels for their students if they want to increase the validity and reliability of their tests.

The findings also revealed that there were some items having high difficulty levels which mean these items were too easy for the students' levels. These easy items do not require high level ability or comprehension to answer them correctly. Hence, these items might lead to inflated scores and a decline in motivation of the students. Students might be misguided by these easy items and they might feel no need to study more. Moreover, these kinds of easy items might include incidental clues and increase the possibility of guessing (Burton, 2005). According to Oluseyi and Olufemi (2012) those items might not be worth even testing. However, Brown, H.D. (2004) suggests that too easy items might not create a big problem for the overall quality of the test if the number of too easy items is limited. On the contrary, too easy items might be benefitted as warm-up activities to increase the motivation especially for low ability students (Coombe, et al., 2007; Gajjar et al., 2014). That way positive washback effect could also be stimulated for low ability students (Alderson & Wall, 1993).

A further argument supporting the availability of easy items is that, if these too easy items are about a very well- known fact and asking basic knowledge on a topic they should not be omitted from the exam (Oppenheim, 2002). However, Haladyna et al. (2002) suggest paraphrasing the language used in the course book or during the instruction to prevent testing for just recall (guideline # 3). Teachers should try to choose a novel material which can be new words such as synonyms even if they target to test older and basic knowledge. In practice, it can be recommended to teachers or test developers to limit the number of easy items, place them at the beginning of the exam and to form the easy items on basic information of topics taught in the class.

Moreover, the results also showed that some items in final exams were too difficult for students. These difficult questions might lead to deflated scores and students' motivation might be declined. They might feel desperate and have the feeling of failure despite all their work and effort. However, difficult items might also be a challenge for high ability students (Brown, H.D., 2004). Nonetheless, test developers or teachers should be cautious to limit the number of too difficult questions to prevent the possible negative washback effect of the exam on test takers. Furthermore, the results highlighted some of the students' difficulties which might help instructors to make changes in their sequence of topics, range of activities, teaching materials or syllabi in their curriculum. Thereby, positive washback effect could be derived from such concerns and changes in teaching materials and methods (Álvarez, 2013).

The primary concern for teachers or test developers should be trying to interpret the item analysis results efficiently. Here, the reasons behind the difficult items play a significant role during the interpretation process. The difficult items might include ambiguous words, high level of language, confusing structure, an incorrect key, or the content of the item might not be clear enough to be understood by the students. However, Haladyna et al. (2002) suggest that direction and content in the stem should be very clear (guideline # 14). In that respect, teachers need to worry about whether they are asking a difficult question or an impossible question to answer (Bodner, 1980). Teachers might be trying to eliminate the surface learners by using difficult language or wording in the stem. However, Burton (2005) claims that writing too complicated items do not help reduce guessing.

Moreover, the difficulty of the questions might also stem from the content. The difficult items might not be based on a siginificant content to learn. The topics that are not emphasized or discoursed during the class might exemplify such kind of topics. However, in their item writing guidelines Haladyna et al. (2002) suggest avoiding trivial content (guideline # 2). These kinds of trifling items may lead to negative washback effect on both high ability and low ability students. Here, test developers or teachers may be inclined to think that difficult items are a must for their exams and they might be nitpicky to ask trivial content. Furthermore, the difficult items might address an over specific or over general content which should also be avoided (Haladyna et al. 2002) (guideline # 5). Moreover, the difficult items might be tricky questions which disadvantage high ability students and decrease their motivation, therefore they should be avoided (Haladyna et al. 2002) (guideline # 7). Moreover, excessive verbalism might prevent the students from answering an item correctly. Therefore, Haladyna et al. (2002) suggest avoiding “window dressing” (guideline # 16, p. 312). Hence, it is obvious that the difficult questions need to be revised and examined to avoid irrelevant difficulties for students.

Research Question 2:What is the discrimination index (item discrimination) of each item on the final exam test administered to non-compulsory preparatory school students?

The findings demonstrated that almost all of the items in final exams had low item discrimination indices. That means these items could not distinguish deep learners and surface learners which is one of the primary aims of assessment. In that case, it is possible to conclude that high ability students were not rewarded on their success while low ability students were not punished. Moreover, it might be maintained that the score of high ability students for that item is not parallel to their score for their overall score on the exam. It is obvious that these items are flawed and they should be edited and proofed for future use (Haladyna et al. (2002) (guideline # 11).

Items having low discrimination indices might not be valid enough for the exam results since ‘discrimination indices’ are also called as ‘validity indices’ (Burton, 2001). One reason behind the poor discriminatory power might be very easy or very difficult questions since those questions tend to discriminate poorly between high ability and low ability students. Another reason might be flawed distractors chosen by high ability students. There should be a negative correlation between examinee’s selection of distractors and their total test scores (DiBattista & Kurzawa, 2011). Therefore, distractors are expected to lure lower ability students rather than higher ability students. However, it may not be the case with items having low discrimination indices.

Moreover, according to the results, some of the items had moderate item discrimination indices. These items might be argued to be more effective and yield more reliable results about the success of the students (Downing, 2005). Moreover, these items contribute to the overall reliability and quality of the exam. When the numbers of items having higher item discrimination indices are increased the exam could be argued to serve to the aims of testing. Items having moderate discrimination index might stimulate positive washback effect for high ability students. These items can be kept and added to the question bank for future use.

The results also showed that, some items had negative discrimination indices. That means these items gathered more correct answers from low ability students rather than high ability students. Thereby, these items with negative discrimination indices detract from the overall quality of a test. The reason behind negative discrimination indices might be a wrong key, two correct answers, or ambiguity in the stem. Still, these items should be modified since they might have a negative backwash effect on high ability students (Hughes, 2003). Hence, these flawed items seemed to penalize high ability students although they were successful (Downing, 2005). Moreover, the success of low ability students on these items might be a result of chance factor or pure guessing (Bush, 2015; Oluseyi & Olufemi, 2012). In that case, it can be recommended to teachers or test developers to revise the stem or the options both the key and the distractors.

It needs to be noted that poorly written items having ambiguity may cause to misunderstanding or different interpretations among the students (Atalmış, 2014). That way, a high ability student might not answer a question while a low ability student can answer it as this is the situation in items having negative discrimination indices. For instance, high ability students might be suspicious about an easy item and they might have regarded the questions as more difficult due to complex wording, structure, or a trick in the stem. Such cases are undesirable for both teachers and high ability students. Therefore, Haladyna et al. (2002) suggest using simple and clear wording in the stem and avoiding tricks (guideline # 14 and guideline # 7). Items with negative discrimination indexes are useless and they decrease the validity of the test. According to Burton (2001) these items should be removed from the test to improve the reliability of the exam.

A solution to increase the discriminatory power or the quality of the items might be revising the items and training test developers and improving the quality of items and distractors (Josefowicz et al. 2002 & Wallach et al. 2006). It would be misleading to assume that all teachers are able to construct well-functioning items without any instruction (Burton, 2005). Item discrimination and plausible distractors are directly related and they are both argued to be the criteria for the quality of a test (Ware & Vik, 2009). The results of item analysis ought to be well analyzed to diagnose who is progressing or who needs extra instruction, which is one of the main objectives of assessment (Coombe, et al., 2007).

Research Question 3: What is the distribution of the response patterns (distractor efficiency) for each of the options on the final exam test administered to non-compulsory preparatory school students?

The statistical tables showed that nearly one third of the items in final exams had at least one non-functioning distractor. These flawed distractors should be edited or modified to be more attractive since they have no utility. No matter how the content is well, structure or wording of the stem, flawed distractors cast a shadow on the quality of the item. Haladyna et al. (2002) introduced 31 item writing guidelines and 14 of them were about writing options both the key (correct answer) and the distractors (incorrect answers). Analyzing the statistical properties of the test items after the test administration is very significant to eliminate the non-functioning distractors from the test for future use. Analyzing each item using item analysis yields significant data to the teachers, and also the institutions for test improvement (Tarrant et al., 2009). Only in that way, "pedagogically and psychometrically sound tests can be developed" (Tarrant et al., 2009, p. 7).

Moreover, the results showed that some distractors could attract more students from high ability students rather than low ability students. These distractors might be even more problematic and they might be discarded or omitted form the test completely. Teachers or test developers might have different reasons to have flawed distractors in their exams. The teachers may not be flexible while writing the distractors since there might be some criteria or rules set by the institution or the exam committee. For instance, the institution might ask the teachers to write five-optioned items. So, the teachers might be trying to find some more option and these options mostly end up being not plausible and written just for the sake of being written (Adisutrisno, 2008). These flawed distractors disadvantage high ability students. However, items on a test do not have to include the same number of options. Some questions might need more options while some others require just one or two distractor because of the content. Ware and Vik (2009) suggest that "whatever number chosen, and this may be quite an arbitrary decision, an important part of quality assurance is to determine that the number of options that function justifies the number set as a policy" (p. 241).

Studies on the number of the options suggest that three choices are adequate (Haladyna et al. 2002; Rogers & Harley, 1999; Bruno & Dirkzwager, 1995; Ebel, 1969; Haladyna & Downing, 1993; Tversky, 1964). In their item writing guidelines Haladyna et al. (2002) suggest decreasing the number of the options (guideline # 18). All the options' being plausible is more important than the number of the options. Writing items with three options might be less time consuming for the teachers. They could spend their energy on writing more items instead of writing more distractors. In the study conducted by Costin (1972) it was found that items having three alternatives had higher discrimination values when compared to the items having four or more alternatives. Similarly, in their study William and Ebel (as cited in Rogers & Harley, 1999) had also reported that two or three-optioned items had equal discrimination indices to four-optioned items. Moreover, in their study Haladyna and Downing (1993) stressed that none of the five-optioned item in their study had four functional distractors. Hence, it seems useless to try to write more options. Instead, it can be recommended to teachers to try to write more items rather than more options. It can enable teachers to cover more content. Furthermore, tests with more items tend to be more reliable than tests with fewer items and more distractors. Furthermore, students' fatigue and test anxiety could be decreased with a shorter test with fewer options. At the same time, students might have more time to read all the questions in the test. Thereby, the reliability of the test can be increased with three options.

Above all, the primary reasons behind flawed and non-functioning distractors might be that the instructors have no training on item writing especially on writing the options.

However, Haladyna et al. (2002) claim that writing the choices of a question is the most difficult part of writing a MC item because the distractors should be plausible and they should base on the common errors of the students. That means a lot of expertise in writing options and experience in knowing your students' mistakes well. Both situations require a great deal of time and effort which might be one of the reasons why some teachers avoid using MC tests for their exams. When the instructors are trained on item writing they would achieve more quality items having strong distractors. That way, more reliable and accurate results about the students' performance could be gathered. Hence, the overall reliability and validity of the exam would be increased.

Designing of good quality tests are very significant since the interpretations of the results affect the learning considerations and outcomes. Hence, an appropriate value of discrimination index, difficulty value and distractor efficiency should be ensured to determine the performance of the students and achievement of the learning objectives (Hamzah & Abdullah, 2011). Teachers should be willing to ensure that their MC exams are of high quality. Institutions should hand item analysis reports to the instructors after each exam administered to the students. Cechova, et al. (2014) notes that after an item analysis, teachers or test developers can decide on what further steps to take in order to increase the reliability or validity of the test. Moreover, analyzing the items could improve the instructors' test construction skills (Oluseyi & Olufemi, 2012). However, if the instructors are not formally trained on developing test items and if they don't even know the terms about item analysis, the reports might not work efficiently. It would be almost impossible interpret the results of the analysis appropriately. Therefore, the instructors who are responsible for preparing the exam should be provided in-service training. This study is in tandem with the findings of Josefowicz et al. (2002), who states that the quality of test items might be significantly improved by providing instructors with formal training on item writing which is a skill that can be learned. As, in this study implications are drawn for test developers and teachers, more rigorous studies of this kind are needed for further research.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115-129. doi: 10.1093/applin/14.2.115.

Adisutrisno, W. D. (2008). Multiple choice English grammar test items that aid English learning for students of English as a foreign language. k@ta, *10*(1), 36-52. Retrieved from https://pdfs.semanticscholar.org/370e/9704ff691a7489fb23012cc7dc57db3d86ff.pdf

Álvarez, I. A. (2013). Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice tests. *International Journal of English Studies, 13*(2), 21-38. doi: 10.6018/ijes.13.2.185861.

Atalmış, H. E. (2014). *The impact of the Test Types and Number of Solution Steps of Multiple-Choice Items on Item Difficulty and Discrimination and Test Reliability.* Published doctoral thesis, University of Kansas, Lawrence, Kansas, USA. Retrieved from: https://pdfs.semanticscholar.org/c18c/cf9bf6aa97437dae5b73a632e6daad65e7b5.pdf

Bodner, G. M. (1980). Statistical Analysis of Multiple-Choice Exams. *Journal of Chemical Education, 57*(3), 188-90. doi: 10.1021/ed057p188.

Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. NY: Pearson Education.

Brown, J. D. (2003). Norm-referenced item analysis (item facility and item discrimination). *Statistics, 7*(2). Retrieved from http://hosted.jalt.org/test/PDF/Brown17.pdf

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959-966. doi:10.1177/0013164495055006004.

Buckles, S., & Siegfried, J. J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education, 37*(1), 48-57. doi: 10.3200/JECE.37.1.48-57.

Burton, R. F. (2001). Do item-discrimination indices really help us to improve our tests? *Assessment & Evaluation in Higher Education, 26*(3), 213-220. doi: 10.1080/02602930120052378.

Burton, R. F. (2004). Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education, 29*(5), 585-595. doi:10.1080/02602930410001689153.

Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72. doi: 10.1080/0260293042003243904.

Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education, 14*(4), 398-404. doi: 10.1108/09684880610703974.

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*(2), 218-231. doi: 10.1080/02602938.2014.902192.

Cechova, I., Neubauer, J., & Sedlacik, M. (2014). Computer-adaptive testing: item analysis and statistics for effective testing. In *European Conference on e-Learning* (p. 106). Academic Conferences International Limited. Retrieved from https://k101.unob.cz/~neubauer/pdf/ECEL2014.pdf

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System, 37*(2), 226-242. doi:10.1016/j.system.2008.11.008.

Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007). *A practical guide to assessing English language learners*. University of Michigan Press.

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*(2), 353-358. doi: 10.1177/001316447003000217.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement, 32*(4), 1035-1038. doi:10.1177/001316447203200419

Çalışkan, H., & Kaşıkçı, Y. (2010). The application of traditional and alternative assessment and evaluation tools by teachers in social studies. *Procedia-Social and Behavioral Sciences, 2*(2), 4152-4156. doi:10.1016/j.sbspro.2010.03.656.

DiBattista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 1-23. doi:10.5206/cjsotl-rcacea.2011.2.4.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82. doi: 10.1207/s15324818ame1001_4.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education, 10*(2), 133-143. doi: 10.1007/s10459-004-4019-5.

Ebel, R. L. (1967). The relationship of item discrimination to test reliability. *Journal of educational Measurement, 4*(3), 125-128. doi:10.1111/j.1745-3984.1967.tb00579.x.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement, 29*(3), 565-570.doi:10.1177/001316446902900302.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine, 39*(1), 17-20. doi:10.4103/0970-0218.126347.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly, 11* (1), 69-78. doi:10.2307/3585593.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education, 2*(1), 37-50. doi:10.1207/s15324818ame0201_3.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999-1010. doi: 10.1177/0013164493053004013.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333. doi:10.1207/S15324818AME1503_5.

Hamzah, M. S. G., & Abdullah, S. K. (2011). Test Item Analysis: An Educator Professionalism Approach. Online Submission. Retrieved from https://files.eric.ed.gov/fulltext/ED524897.pdf

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Pres.

Jafarpur, A. (1999). Can the C-test be improved with the classical item analysis? *System, 27*(1), 79-89. doi:10.1016/S0346-251X(98)00043-8.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine, 77*(2), 156-161. doi:10.1097/00001888-200202000-00016.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education, 37*(8), 919-931. doi: 10.1080/02602938.2011.586991.

Oluseyi, A. E., & Olufemi, A. T. (2012). The Analysis of Multiple Choice Item of the Test of an Introductory Course in Chemistry in a Nigerian University. *International Journal of Learning, 18*(4), 237-246. doi:10.18848/1447-9494/CGP/v18i04/47579.

Oppenheim, N. (2002). Empirical analysis of an examination based on the academy of legal studies in business test bank. *Journal of Legal Studies Education, 20*(2), 129-158. doi: 10.1111/j.1744-1722.2002.tb00135.x.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234-247. doi:10.1177/00131649921969820.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286. doi:10.1111/j.1745-3984.1985.tb01064.x.

Tarrant, M., Ware, J. & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*(1), 40. doi:10.1186/1472-6920-9-40.

Temizkan, M., & Sallabaş, M. E. (2011). Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması [English translation here nedede]. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 30*(issue), 207-220. Retrieved from https://dergipark.org.tr/tr/download/article-file/55711

Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies, 8*(6), 141-146. doi:10.7575/aiac.alls.v.8n.6p.140.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*(2), 386-391. doi:10.1016/0022-2496(64)90010-0.

Üstüner, A., & Şengül, M. (2004). Çoktan seçmeli test tekniğinin Türkçe öğretimine olumsuz etkileri [Negative Effect of Multiple ChoiceTest Technic on Turkish Teaching]. *Fırat Üniversitesi Sosyal Bilimler Dergisi, 14*(*2*), 197-208. Retrieved from http://web.firat.edu.tr/sosyalbil/dergi/arsiv/cilt14/sayi2/197-202.pdf

Wainer, H. (1988). The future of item analysis. *ETS Research Report Series,* 1988(2). Retrieved fromhttps://www.jstor.org/stable/pdf/1434865.pdf?refreqid=excelsior%3Aef5d058bdbca6cbe279cb67311b2612f

Wallach, P. M., Crespo, L. M., Holtzman, K. Z., Galbraith, R. M., & Swanson, D. B. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education, 11*(1), 61-68. doi:10.1007/s10459-004-7515-8.

Ware, J., & Vik, T. (2009). Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Medical teacher, 31*(3), 238-243. doi:10.1080/01421590802155597.

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kulanilan örnekleri [Multiple choice item types and samples used in science education]. . *Gazi Eğitim Bilimleri Dergisi, 2*(2), 151-170. Retrieved from https://dergipark.org.tr/tr/download/article-file/419772

Yıldırım, O. (2010). Washback effects of a high-stakes university entrance exam: Effects of the English section of the university entrance exam on future English language teachers in Turkey. *The Asian EFL Journal Quarterly, 12*(2), 92-116. Retrieved from https://asian-efl-journal.com/PDF/June-2010.pdf