

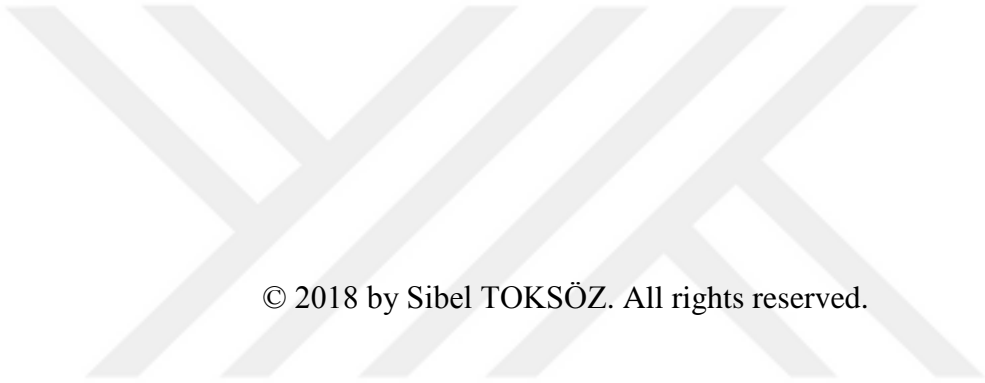
**T.C.**  
**SÜLEYMAN DEMİREL UNIVERSITY**  
**GRADUATE SCHOOL OF EDUCATIONAL SCIENCES**  
**DEPARTMENT OF FOREIGN LANGUAGE EDUCATION**

**ITEM ANALYSIS OF MULTIPLE CHOICE FINAL EXAMS OF NON-  
COMPULSORY PREPARATORY STUDENTS: A STUDY AT A STATE  
UNIVERSITY**

**Sibel TOKSÖZ**

**Advisor: Assoc. Prof. Nazlı BAYKAL**

**MASTER'S THESIS**  
**ISPARTA 2018**




© 2018 by Sibel TOKSÖZ. All rights reserved.

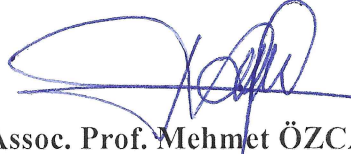
## CERTIFICATE OF COMMITTEE APPROVAL

We certify that this thesis under the title of “**Item Analysis of Multiple Choice Final Exams of Non-Compulsory Preparatory Students: A Study at a State University**” prepared by **Sibel TOKSÖZ** is satisfactory for the award of the degree of Master of Arts in the Department of Foreign Language Education.

Advisor

  
Assoc. Prof. Nazlı BAYKAL  
Süleyman Demirel University

Committee Member

  
Assoc. Prof. Mehmet ÖZCAN  
Mehmet Akif Ersoy University

Committee Member

  
Assist. Prof. Kağan BÜYÜKKARCI  
Süleyman Demirel University

Director

  
Prof. Seyfettin ÇAKMAK

## COMMITMENT

I declare that this thesis has been written by taking ethical rules into consideration and by giving all the references cited from the field by referring them in the thesis.



**Sibel TOKSÖZ**

## TABLE OF CONTENTS

|  |      |
|--|------|
| TABLE OF CONTENTS.....                           | i    |
| ABSTRACT.....                                    | iv   |
| ÖZET .....                                       | vi   |
| ACKNOWLEDGEMENTS.....                            | viii |
| LIST OF TABLES.....                              | ix   |
| LIST OF ABBREVIATIONS.....                       | xi   |
| 1. INTRODUCTION .....                            | 1    |
| 1.1. Statement of the Problem.....               | 1    |
| 1.2. Purpose of the Study .....                  | 3    |
| 1.3. Significance of the Study .....             | 4    |
| 1.4. Assumptions.....                            | 5    |
| 1.5. Limitations .....                           | 6    |
| 2. CONCEPTUAL FRAMEWORK AND RELATED STUDIES..... | 7    |
| 2.1. Assesment in English Language Teaching..... | 7    |
| 2.2. Multiple Choice Tests.....                  | 8    |
| 2.2.1. Advantages of multiple choice tests ..... | 10   |
| 2.2.2. Criticism on multiple choice tests.....   | 13   |
| 2.3. Item Analysis .....                         | 17   |
| 2.3.1. Item facility .....                       | 19   |
| 2.3.1.1. Studies on item facility .....          | 21   |
| 2.3.2. Item discrimination .....                 | 22   |
| 2.3.2.1. Studies on item discrimination .....    | 26   |
| 2.3.3. Distractor efficiency .....               | 28   |
| 2.3.3.1. Studies on distractor efficiency.....   | 31   |
| 3. METHODOLOGY .....                             | 34   |
| 3.1. Research Design .....                       | 34   |
| 3.2. Participants.....                           | 34   |
| 3.3. Data Collection Process .....               | 35   |
| 3.4. Data Collection Tools .....                 | 36   |
| 3.4.1. Final exams .....                         | 36   |
| 3.4.2. Semi-structured interview .....           | 37   |
| 3.5. Data Analysis.....                          | 38   |

|   |    |
|---|----|
| 3.5.1. Quantitative data analysis .....   | 38 |
| 3.5.2. Qualitative data analysis .....  | 39 |
| 4. RESULTS .....  | 40 |
| 4.1. Quantitative results of the item analysis of the multiple choice final exams .....   | 40 |
| 4.1.1. Item facility indices of the items on the final test exams.....  | 40 |
| 4.1.2. Item discrimination indices of the items on the final test exams.....  | 43 |
| 4.1.3. Distractor efficiency of the options of the items on the final test exams.....   | 47 |
| 4.1.4. The success of the students according to the parts of the exam .....   | 52 |
| 4.1.5. Non-compulsory preparatory school students' in-year grades point average and the students' final grades point average.....   | 53 |
| 4.2. Non-compulsory preparatory school students' perceptions about the difficulty level of the items, the discriminatory power of the exam, the efficiency of the distractors and the exam in general .....                   | 54 |
| 4.2.1. Participants' remarks about the difficulty level of the items in the final exam ...  | 54 |
| 4.2.2. Participants' remarks about the discriminatory power of the items in the final exam.....   | 55 |
| 4.2.3. Participants' remarks about the efficiency of the distractors .....  | 55 |
| 4.2.4. Participants' remarks about their feelings and motivations when they were torn between two options in the final exam .....   | 56 |
| 4.2.5. Participants' remarks about the most difficult and easiest part of the final exam  | 57 |
| 4.2.6. Participants' final words about the final exam .....   | 59 |
| 5. DISCUSSION AND CONCLUSION .....  | 60 |
| 5.1. What is the difficulty level (item facility) of each item on the final exam test administered to non-compulsory preparatory school students at MAKU?.....  | 60 |
| 5.2. What is the discrimination index (item discrimination) of each item on the final exam test administered to non-compulsory preparatory school students at MAKU? ....  | 63 |
| 5.3. What is the distribution of the response patterns (distractor efficiency) for each of the five options of the items on the final exam test administered to non-compulsory preparatory school students at MAKU like?..... | 66 |
| 5.4. In which part of the exam did the students do well, and in which parts did the students do badly?.....   | 69 |
| 5.5. Is there a significant relationship between the non-compulsory preparatory school students' in-year grades point average and the students' final grades point average? ...   | 70 |

|  |     |
|--|-----|
| 5.6. What are the non-compulsory preparatory school students' perceptions about the difficulty level of the items, the discriminatory power of the exam, the efficiency of the distractors and the exam in general administered to the non-compulsory preparatory school students at MAKU? ..... | 70  |
| 5.7. Conclusion .....  | 71  |
| REFERENCES .....   | 74  |
| APPENDICES .....   | 80  |
| Appendix A. Final Exam (Session I) Used in the Study .....   | 81  |
| Appendix B. Final Exam (Session II) Used in the Study .....  | 90  |
| Appendix C. Interview (Turkish) Used in the Study .....  | 99  |
| Appendix D. Interview (English) Used in the Study .....  | 100 |
| Appendix E. Permission Document.....   | 101 |
| CURRICULUM VITAE.....  | 102 |

## **ABSTRACT**

### **ITEM ANALYSIS OF MULTIPLE CHOICE FINAL EXAM OF NON-COMPULSORY PREPARATORY STUDENTS : A STUDY AT A STATE UNIVERSITY** **Sibel TOKSÖZ**

**Master's Thesis, Süleyman Demirel University, Graduate School of Educational Sciences, Department of Foreign Language Education**

**Advisor: Assoc. Prof. Nazlı BAYKAL**

**2018, 102 pages**

The aim of this study is to examine the multiple choice final exams administered to 210 non-compulsory preparatory school students at Mehmet Akif Ersoy University (henceforward MAKU). The study specifically aims to analyze the exams in terms of three characteristics of multiple-choice questions (MCQs): item facility, item discrimination and distractor efficiency. Although there have been many research papers studying multiple choice tests, there have been very limited studies analyzing the multiple choice items in terms of item analysis in Turkish literature. Hence, this study will contribute to the field analyzing two multiple-choice exams with respect to item analysis. In this study mixed-method research design was adopted which benefitted from statistical analysis and a semi-structured interview. The data were analyzed through Paired Samples T-Test, Frequency analysis and content 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 showed that one third of the items in the exams had at least one non-functional distractor. The results showed that the students in session 1 did best in Listening part while they did worst in Dialogue part. However, the students in session 2 did best in Translation part while they did worst in Vocabulary part. Moreover, there was found a significant difference between the students' in-year grades point average and final grades point average. Finally, the results showed that one third of the participants stated that the exam was very difficult for their levels while one third of the participants remarked that the exam had a moderate difficulty level, and the rest one third of them stated that the exam was



easy. Moreover, most participants remarked that the exam was discriminating well and the distractors were efficient. At the end of the study, some guidelines were presented for teachers and test developers to make the items more functional having appropriate mix of difficulty levels with a high discrimination and effective distractors.

**Keywords:** Item Analysis, Multiple Choice, Item Facility, Item Discrimination, Distractor Efficiency



## ÖZET

# İSTEĞE BAĞLI HAZIRLIK ÖĞRENCİLERİNE UYGULANAN ÇOKTAN SEÇMELİ FİNAL SINAVININ MADDE ANALİZİ: BİR DEVLET ÜNİVERSİTESİ'NDE YAPILMIŞ BİR ÇALIŞMA

Sibel TOKSÖZ

**Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Eğitim Bilimleri Enstitüsü,  
Yabancı Diller Eğitimi Anabilim Dalı  
Danışman: Doç. Dr. Nazlı BAYKAL  
2018, 102 sayfa**

Bu çalışmanın amacı Mehmet Akif Ersoy Üniversitesinde (MAKU) isteğe bağlı hazırlık okuyan 210 öğrenciye uygulanan çoktan seçmeli final sınavlarını incelemektir. Bu çalışma sınavları özellikle çoktan seçmeli soruların üç niteliği yani madde zorluk derecesi, madde ayırt ediciliği 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 yarı yapılandırılmış görüşmeden ve istatistiksel analizlerden yararlanan karma araştırma yöntemi kullanılmıştır. Verileri analiz etmek için tek örneklem t-testi, sıklık analizi ve içerik analizi kullanılmıştır. Sonuçlar göstermektedir ki, sınavlardaki soruların çoğu öğrenciler için orta zorluk derecesine sahiptir. Ancak, soruların hemen hepsi çok düşük madde ayırcılığı değerine sahiptir ve bazı soruların ayırcılık değerinin negatif olduğu bulunmuştur. Ayrıca, sonuçlar sınavlardaki soruların üçte birinde en az bir tane işlevsiz çeldirici olduğunu göstermektedir. Sonuçlara göre 1. oturumdaki öğrencilerin en başarılı olduğu bölüm Dinleme iken, en başarısız oldukları bölüm ise Diyalog bölümüdür. 2. oturumdaki öğrencilerin en başarılı oldukları bölüm Çeviri bölümü iken en başarısız oldukları bölüm Kelime bölümüdür. Ayrıca, öğrencilerin yıl içi sınav not ortalamaları ile final sınavı not ortalamaları arasında anlamlı bir fark olduğu bulunmuştur. Son olarak, sonuçlar göstermektedir ki katılımcıların üçte biri sınavın çok zor olduğunu düşünürken üçte biri sınavın orta zorlukta olduğunu ve geri kalan üçte biri ise sınavın

kolay olduđunu belirtmiřtir. Ayrıca, katılımcıların çođu sınavın iyi bir ayırt ediciliđe sahip olduđunu ve eldiricilerin etkili olduđunu belirtmiřtir. alıřmanın sonunda, soruları ideal ve farklı kolaylık derecesine, yüksek ayırt ediciliđe ve etkili eldiricilere sahip olacak řekilde daha etkili hale getirmek iin retmenler ve soru geliřtirenler iin bazı ynergeler sunulmuřtur.

**Anahtar Kelimeler:** Madde Analizi, oktan Semeli, Madde Zorluk Derecesi, Madde Ayırt Ediciliđi, eldirici Yeteneđi



## ACKNOWLEDGEMENTS

This study wouldn't have been completed without the support and guidance of some people and I wish to express my thanks to them.

First of all, I would like to express my most sincere gratitude and appreciation to my advisor Assoc. Prof. Nazlı BAYKAL for her valuable support, guidance, generous motivation, partnership and understanding during the process of writing this thesis. I am also very much indebted to Assist. Prof. Kağan BÜYÜKKARCI for his professional advice, generous help, and irreplaceable comments in many steps of this thesis.

I also owe special thanks to Assoc. Prof. Mehmet ÖZCAN for his inspiring suggestions, invaluable comments, and enthusiasm. And I am also thankful to Assoc. Prof. Mustafa ŞEVİK who collaborated in collecting data and gave me support. I also owe thanks to Assoc. Prof. Ferit KILIÇKAYA who encouraged me and helped me with his valuable guidance.

I owe endless thanks to my family, my father Remzi BAYIR and my mother Meziyet BAYIR who have given me enormous amount of support with their sincere love, unending care, and patience throughout my whole educational life. They have always been with me there although they are far away. I am also thankful to my sisters and brothers for being such wonderful siblings and having priceless smiles.

It is also a debt for me to thank everybody who collaborated and helped me to complete this thesis especially my master friends for their support and help and my preparatory students who participated voluntarily in my research by giving their time.

Finally, my deepest thank goes to my beloved husband İsmet TOKSÖZ for his eternal love, endless patience and belief in me and my capacity. He has always been my pole star and light whenever I am lost. I feel very lucky for having such a soul and life mate in my life.

## LIST OF TABLES

|  |    |
|--|----|
| Table 1. General content of the final exams administered to preparatory students.....    | 36 |
| Table 2. Item Facility (IF) indices of too easy items in final exam session 1 .....      | 40 |
| Table 3. Item Facility (IF) indices of too difficult items in final exam session 1.....  | 40 |
| Table 4. Item Facility (IF) indices of moderate difficult items in final exam session 1. | 41 |
| Table 5. Item Facility (IF) indices of too easy items in final exam session 2 .....      | 41 |
| Table 6. Item Facility (IF) indices of moderate difficult items in final exam session 2. | 42 |
| Table 7. Item Facility (IF) indices of moderate difficult items in final exam session 2. | 42 |
| Table 8. Overall Item Facility (IF) results of final exams .....                         | 43 |
| Table 9. Items with negative discrimination indices in final exam session 1.....         | 43 |
| Table 10. Distributions of responses for item # 60 in final exam session 1 .....         | 43 |
| Table 11. Items with moderate discrimination indices in final exam session 1 .....       | 44 |
| Table 12. Items with zero or low discrimination indexes in final exam session 1 .....    | 44 |
| Table 13. Items with negative discrimination indices in final exam session 2.....        | 45 |
| Table 14. Distribution of responses for item # 6 in final exam session 2.....            | 45 |
| Table 15. Distribution of responses for item # 17 in final exam session 2.....           | 45 |
| Table 16. Distribution of responses for item # 57 in final exam session 2.....           | 46 |
| Table 17. Items with zero or low discrimination indexes in final exam session 2 .....    | 46 |
| Table 18. Items with moderate discrimination indices in final exam session 2 .....       | 47 |
| Table 19. Overall Discrimination Index (DI) results of final exams .....                 | 47 |
| Table 20. Items with NFDs in Listening part of final exam session 1 .....                | 48 |
| Table 21. Items with NFDs in Grammar part of final exam session 1 .....                  | 48 |
| Table 22. Items with NFDs in Vocabulary part of final exam session 1 .....               | 48 |
| Table 23. Items with NFDs in Cloze Test part of final exam session 1 .....               | 49 |
| Table 24. Items with NFDs in Translation part of final exam session 1 .....              | 49 |
| Table 25. Items with NFDs in Reading part of final exam session 1 .....                  | 49 |
| Table 26. Items with NFDs in Listening part of final exam session 2 .....                | 50 |
| Table 27. Items with NFDs in Grammar part of final exam session 2 .....                  | 50 |
| Table 28. Items with NFDs in Vocabulary part of final exam session 2.....                | 51 |
| Table 29. Items with NFDs in Cloze Test part of final exam session 2 .....               | 51 |
| Table 30. Items with NFDs in Reading part of final exam session 2 .....                  | 51 |
| Table 31. Items with NFDs in Translation part of final exam session 2 .....              | 51 |
| Table 32. Items with NFDs in Dialogue part of final exam session 2.....                  | 52 |

Table 33. The success of the students according to the parts of the final exam session 1  
..... 52

Table 34. The success of the students according to the parts of the final exam session 2  
..... 53

Table 35. T-test results for students' in-year grades and final grades point average ..... 53



## LIST OF ABBREVIATIONS

|       |                               |
|-------|-------------------------------|
| MC    | Multiple-Choice               |
| MAKU  | Mehmet Akif Ersoy University  |
| MCQs  | Multiple Choice Questions     |
| ELT   | English Language Teaching     |
| EFL   | English as a Foreign Language |
| ESL   | English as a Second Language  |
| DIF I | Difficulty Index              |
| DI    | Discrimination Index          |
| DE    | Distractor Efficiency         |
| IF    | Item Facility                 |
| NOTA  | None of the Above             |
| NFD   | Non- functional Distractors   |
| QUAN  | Quantitative                  |
| QUAL  | Qualitative                   |
| ID    | Item Discrimination           |
| P     | Participant                   |

## **1. INTRODUCTION**

This chapter presents statement of the problem, purpose of the study, research questions and significance of the study, as well as the assumptions and limitations. Each section is aimed to enable a better understanding and coverage of the study.

### **1.1. Statement of the Problem**

As a foreign language, English has been very popular and also a problematic issue for years in Turkey. Apart from the academic privileges, English has been a gate for many opportunities such as obtaining a profession, getting a higher salary or being promoted. Therefore, there has been a growing recognition of the vital links between the success in English and those privileges. 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).

A typical MC item consists of a question which is also referred to as stem, a correct option which is the key, and two or more other options which are called distractors. However, MC tests may vary in terms of their length, syntactic complexity, level of vocabulary, and topical content (Bachman, 1991). In an MC test, the student is supposed to choose the best option that is the key or answer of the question posed. Teachers prefer MC tests to examine efficiently large numbers of students on a broad variety of topics in one exam and to provide quicker feedback compared to other forms of traditional assessment tools (Bush, 2001; Williams & Clark, 2004; Simkin & Kuechler, 2005; Nicol, 2007). Besides teachers, MC tests are also favored by students because they think that MC tests are objective and they can get points even if they cannot answer all of the questions on the test (Simkin & Kuechler, 2005).

In addition to exams at schools, the overall proficiency in English has been determined according to the results of some standardized tests administered by Turkish Republic Assessment, Selection and Placement Center (ÖSYM) such as YDS or e-YDS. These tests are high-stakes and both are in multiple-choice format. Therefore, multiple choice tests have been much more in demand because of the gate keeping roles of those exams.



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" (Olufemi & Oluseyi, 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 (Olufemi & Oluseyi, 2012). With the help of item analysis, too easy or too difficult items can be identified and they can be dropped or at 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).

Item facility also referred to as item difficulty, is defined as the extent to which an item is easy or difficult for a determined group of test takers (Brown, 2004). Item facility is a crucial part of item analysis. Jafarpur (1999) stresses the need for item facility calculations stating that in many tests "some lexical items are either very easy to predict or extremely hard to guess" (p.80). Another concept is item discrimination which refers to the extent to which an item differentiates between high and low ability test takers (Brown, 2004). Item discrimination has an important role in the reliability of a test. Ebel (1967) points out that discriminatory power is a chief determinant of the quality of an MC item. Moreover, Goodrich (1977) discusses that "an effective question and each distractor should have a degree of potency and discrimination" (p.70). The last term, distractor efficiency is about how the responses are distributed to the distractors. Goodrich (1977) notes that a distractor's efficiency can be determined according to its ability, to separate the students whose proficiency levels are different. Hence, choosing the right distractors is a significant task while constructing MC items.

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, advantages or disadvantages of MC tests. Too little attention has been paid to MC tests in terms of item analysis. Hence, it seems prudent to examine the quality of the MC items, which is the primary aim of the present study. In the light of these issues, the study was carried out in order to

investigate whether the final exams of the preparatory students at Mehmet Akif Ersoy University (MAKU) require the necessary qualifications in terms of item analysis or not. The study was also supposed to bridge the gap in the lack of studies about MC tests and item analysis in Turkey. The study questions were developed by the researcher and will be presented in the next section.

## **1.2. Purpose of the Study**

The objective of this study is to examine the multiple choice final exam aiming to test grammar, vocabulary, listening and reading comprehension and administered to the non-compulsory preparatory school students at Mehmet Akif Ersoy University (MAKU). The reason for choosing the final exam is that the final exam has an important role and weight in the preparatory students' overall achievement for the whole academic year; it affects 50 percent of the overall score of the students. All the classes in preparatory school take this same exam although they take different written quizzes or midterms by their instructors. 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. With this respect, the present study aims to investigate the difficulty level of the items, analyzing how many students chose the correct answer for each item. The study also seeks to analyze the items in terms of item discrimination, investigating whether the items discriminated between the high-achieving and the low-achieving students. Moreover, the study aims to analyze the items' distractor efficiency, analyzing the degree of potency of the options. Besides, the study intends to find out in which parts of the study the students did well and in which parts they performed poorly. The study also tries to find out a relationship between the students' in-year grades point average and the students' final grades point average. Finally, the study targets to explore the test takers' feelings and ideas about the test items, the options, and the exam in general, to find out if there is a discrepancy or consistency with the quantitative analysis results. 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 at MAKU?
2. What is the discrimination index (item discrimination) of each item on the final exam test administered to non-compulsory preparatory school students at MAKU?

3. What is the distribution of the response patterns (distractor efficiency) for each of the five options of the items on the final exam test administered to non-compulsory preparatory school students at MAKU like?
4. In which part of the exam did the students do well, and in which parts did the students do badly?
5. Is there a significant relationship between the non-compulsory preparatory school students' in-year grades point average and the students' final grades point average?
6. What are the non-compulsory preparatory school students' perceptions about the difficulty level of the items, discriminatory power of the exam, the efficiency of the distractors and the exam in general administered to the non-compulsory preparatory school students at MAKU?

### **1.3. Significance of the Study**

Multiple choice tests are preferred by most of the teachers or institutions for a variety of disciplines in Turkey. As a foreign language, the students' success or general proficiency in English is mostly tested by multiple choice exams (Goodrich, 1977). Therefore, the results of the study are supposed make an important contribution to English Language Teaching (ELT) as DiBattista and Kurzawa (2011) state many multiple choice tests need to be improved. DiBattista and Kurzawa (2011) suggest the instructors "to consider improving the quality of their multiple-choice tests by conducting an item analysis and by modifying distractors that impair the discriminatory power of items" (p. 1). Hence, the present study will contribute to Turkish literature by analyzing the items of a multiple choice format final test to find out whether the items are functioning correctly or not.

This study also aims to give guidelines or suggestions to modify the items having low discriminating power between the high achieving and low achieving students for future use of the institution as Burton (2001) maintains that indices, especially item-total correlations should be calculated "to use them as guides when reconsidering in detail the content and wording of individual test items" (p. 219). Related to the aims of the study, Jafarpur (1999) points out that test analysis should include the calculation of facility and discrimination of the items. For that reason, the study aims to calculate the

facility and discrimination indices of the items besides distractor efficiency to find out whether the test can be said to be reliable and valid.

Also, Goodrich (1977) points out that although there have been many studies about tests and measurement, few studies seem to have been undertaken to investigate the area of alternate choices. With this respect, this study also aims to find out the efficiency of the distractors since previous studies indicate a need for this kind of studies.

Although there have been many research papers studying multiple choice tests, to the best knowledge of the researcher, there have been very limited studies analyzing the multiple choice items in terms of item analysis. Rodgers and Harley (1999) state that more empirical studies in a variety of subject areas should be done on item analysis. Therefore, this study is supposed to make a major contribution to the field by analyzing the multiple choice format final exam test administered to the non-compulsory preparatory students at MAKU.

Furthermore, this study is expected to yield significant results for MAKU. The students are expected to get a “Certificate of Achievement” at the end of the academic year and the final exam affects the 50 percent of the overall score of the students. Therefore, to analyze the exam in terms of item characteristics appears to be very significant. Since the final exam questions at MAKU School of Foreign Languages are written, assembled and reviewed by the instructors, the findings of the present study have the potential to revise or modify and exclude or keep the items for future use. With this respect, this study is believed to provide guidelines or suggestions about how to modify the items in the exam for future use.

Finally, the study is expected to produce useful results not only for MAKU but also for other undergraduate universities in Turkey as multiple-choice tests are widely used to assess the achievement in second language in most of the state and private universities.

#### **1.4. Assumptions**

The results of the final exam are the quantitative data of this study. Therefore, it is assumed that the students did not cheat during the exam. Also, the interview which is

the qualitative data instrument of the study is assumed to include all the necessary questions to aid the researcher in the data collection process. It is also assumed that there is no ambiguity in the interview questions. Moreover, the students participating in the interview are assumed to answer the questions sincerely and honestly.

### **1.5. Limitations**

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. Another potential limitation of the study might be that the participants of the study are non-compulsory preparatory school students at MAKU. Hence, the results might be influenced by the characteristics of these students at non-compulsory preparatory school at MAKU.

The qualitative data was collected from 21 (10 percent of the all participants) participants. More participants could have been reached to get more generalizable results for the larger groups. Also, the quantitative data was obtained from the final exam only. The study could have enabled a broader picture analyzing the other exams administered during the whole year such as quizzes and midterms.

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.

## 2. CONCEPTUAL FRAMEWORK AND RELATED STUDIES

This chapter will present the background theoretical information and related empirical studies related to multiple choice tests and item analysis namely item facility, item discrimination, and distractor efficiency.

### 2.1. Assessment in English Language Teaching

Getting a grade or a certificate proving that someone is proficient enough in English brings many privileges and opportunities to that person such as being promoted, getting a higher salary, being accepted to a position, or pursuing an academic career. Hence, assessing language proficiency has been an important concept in the studies of English Language Teaching (ELT) (Babaii & Ansary, 2001). Assessment might be defined as a way to gather information about the learners' success or skills with the help of many sorts of tools (Coombe et al., 2007). Throughout their teaching practice, language teachers have ultimately assessed their students' language skills in one way or another (Brown & Hudson, 1998). In that respect assessment makes a substantial contribution to the well-being of teaching and learning practice.

The concepts of *testing* and *assessment* are sometimes used interchangeably although they are different. Test can be defined as "a method of measuring a person's ability, knowledge or performance in a given domain" while assessment is a wider concept defining an "ongoing process" in teaching and learning (Brown, 2004, p. 3). The varieties of language assessment might be classified mainly as selected-response assessments, constructed-response assessments and personal-response assessment. Selected-response assessments include true-false, matching and multiple choice; constructed-response assessments include fill-in, short-answer, and performance; personal-response assessments include conference, portfolio, and self- or peer assessment (Brown & Hudson, 1998).

Each assessment method with a different acronym has been claimed to be better than the previous ones and be able to assess higher order thinking skills (Schuwirt & Van Der Vleuten, 2004). However, among the different forms of assessment, testing has been "the most prevalent in ELT all over the world in the past 50 or more years" (Leung & Lewkowicz, 2006, p.212). Despite ample procedures or tasks to assess students, teachers use tests inevitably (Brown, 2004) although trying to choose a mere assessment

tool may hinder the teachers from assessing some of the topics and neglect them completely since they are not suitable for that specific format (Schuwirt & Van Der Vleuten, 2004).

## **2.2. Multiple Choice Tests**

Among the other testing types, multiple choice (MC) tests have been a major concern in EFL (English as a foreign language) or ESL (English as a second language) contexts. New trends in technology, decreasing resources, and increasing number of students led to the growth in usage of MC tests as an assessment tool in higher education (Nicol, 2007). Simkin and Kuechler (2005) state that although there are many types of assessment tools, most of the instructors and students prefer MC tests for different reasons. Similarly, Coombe et al. (2007) maintain that teachers benefit from tests to assess their students irrespective of their experience in developing tests.

As selected-response assessment tools, in MC tests, students are supposed to choose the best option that is the key or answer of the question posed. Therefore, multiple choice tests are sometimes called as ‘single best-answer multiple choice’ (Tarrant et al., 2009). Apart from the key, there are some other options which are called “distractors”. Distractors are supposed to confuse students if they are not sure about the correct choice. However, weak distractors are unable to trap the students with high knowledge.

MC tests have been widely used to test students’ achievement or general proficiency in English for all grades (Bush, 2001; Bush, 2015; Goodrich, 1977; Schuwirt & Van Der Vleuten, 2004). Multiple choice questions (MCQs) are possibly the most commonly preferred type in the professionally developed tests, in textbooks and English language proficiency exams (Coombe et al., 2007). They are preferred both in classroom environments with small or large enrollments and as stakeholders to enter a university, to get a job or to get a certificate of proficiency in English. Bush (2015) maintains that for low-stake exams, traditional multiple choice tests may have a priority since they are simple and familiar. In addition to low stake exams, MC tests are extensively used on online exams, entrance exams, and certification tests (Simkin & Kuechler, 2005). Furthermore, MC tests are extensively preferred for high-stake exams and that may also enhance their popularity (Downing, 2002).

Test banks offered by different publishers made MC questions even more appealing for both teachers and students (Yonker, 2011). Especially, for grammar and vocabulary questions MC tests are preferred mostly by teachers as Álvarez (2013) pointed out, “today, having completed all the necessary stages of validation process the decision to use a grammar and vocabulary MC tests seems to have been a good solution” (p.23). Besides grammar and vocabulary, multiple choice tests are also preferred by reading experts for assessing reading subskills and distinguishing the main and specific ideas referred in the texts (Coombe et al., 2007). Testing boards argue that comprehension should go beyond sentence level; hence, reference, discourse markers, and insertion points for missing sentences could all be assessed in a single MC test (Coombe et al., 2007). Because of its practical and familiar format, multiple choice tests are also preferred to assess listening skills (Yanagawa & Green, 2008). Similarly, Brown and Hudson (1998) argue that using MC tests to assess reading, listening, grammar knowledge, and phoneme discrimination provides the teachers efficient insights about their students’ abilities.

Being already baffled by intense curriculum, heavy schedule, and large enrollments in class, multiple choice tests might be a good option for teachers to assess what the students can do with their knowledge (Ware & Vik, 2009). Most teachers might be inclined to think that they can assess their students better with essay type exams; Walstad and Becker (1994) claim that some essay type questions do not have much thing to contribute to the results of a well written MC test questions. Well written MC items can measure students’ progress in a broad variety of content accurately (Hansen & Dexter, 1997). Recent improvements in MC tests formats such as item analysis also increased the popularity of the format (Bacon, 2003). Item analysis might help teachers to achieve more functional and well- constructed MC items in addition to making test construction less arduous and time-consuming.

Given the above concerns it seems clear that MC tests are mostly used in EFL classroom context and their significance “seems likely to grow further with the advent of e-learning” (Bush, 2006, p.398). Although the literature seems to be supportive of MC tests, it should not mean to abandon other assessment tools (Bacon, 2003). The reason for preferences and the popularity of MC tests as assessment tools might be attributed to some advantages they provide to teachers and administrators.



### **2.2.1. Advantages of multiple choice tests**

Multiple choice tests have been such popular and preferable because of some advantages they have by nature. Among their advantages it could be argued that MC tests supply a better coverage of the content (Simkin & Kuechler, 2005; Bacon, 2003), measuring a wide variety of topics (Brown & Hudson, 1998; Walstad & Becker, 1994), with efficient format (Bacon, 2003; Walstad & Becker, 1994), and minimal assessor's bias (McCoubrie, 2004; Gajjar et al., 2014), as well as being scored easily in a quick and objective way (Brown & Hudson, 1998; DiBattista & Kurzawa, 2011; Gajjar et al., 2014; Rogers & Harley, 1999), thereby leading to an increased reliability (Rogers & Harley, 1999). Moreover, MC tests could be efficiently used in different educational levels from elementary level to advanced or graduate-levels (Coombe et al., 2007). Being a cost effective assessment tool, multiple choice tests are also favored by administrators in addition to teachers and test designers (Coombe et al., 2007).

Besides their fairness and low cost in grading, MC tests are also favored for the availableness of the statistical analysis they provide to the researchers (Buckles & Siegfried, 2006). Bodner (1980) states another advantage of multiple choice tests is "the ability to calculate a variety of data which pertain to the quality, or perhaps the reliability, of the exam, the extent to which the exam discriminates between "good" and "poor" students"(p. 189). Moreover, MC tests are suitable for e- learning since they are graded easily and automated (Bush, 2006).

Multiple choice tests are also favored for their objectivity in scoring (Brown & Hudson, 1998). In other assessment types such as essays, the same answer can be scored differently by different raters. The students also prefer MC tests since they think it prevents instructor's bias (Simkin & Kuechler, 2005). Contrary to other assessment tools having open ended questions or tasks, in MC tests the raters are objective and unbiased to the test takers since there is only a single key answer for each question (Atalmış, 2014). As well as objectivity, multiple choice tests offer consistency in grading (Bodner, 1980). Because of the format of MC tests the grades are ensured to be consistent across the board. In addition, Bacon (2003) maintains that "MC tests can yield scores at least as reliable as those produced by a constructed-response test, while also allowing for broader coverage of the topics covered in a course" (p.1). Simkin and

Kuechler (2005) state that, when MC tests and constructed-response tests assess the same-level content, students get parallel or similar scores.

Easiness of preparing many versions of the same questions to prevent the students' cheating represents another upside of MC tests (Simkin & Kuechler, 2005). For an open-ended format exam, preparing alternative exam papers could be a burden for the instructors having a loaded course schedule. Besides controlling cheating, MC tests are also preferred because of its easiness and fastness in grading and thereby providing a quick feedback to the students about their performances on the test (Nicol, 2007; Simkin & Kuechler, 2005). Examining the incorrect answers on the test, both teachers and students could obtain diagnostic feedback (Hansen & Dexter, 1997). However, the feedback could be very limited and may not be organized according to the needs of the students individually (Nicol, 2007).

Buckles and Siegfried (2006) argue that MC questions can assess in-depth understanding to some extent while they fail to measure it fully in a variety of subjects. Similarly, Bush (2001) argues that in addition to factual knowledge, MC exams can test comparatively high level skills such as cognitive and analytical ones as well as higher level of cognitive reasoning (Schuwirth & Van Der Vleuten, 2004). Coombe et al. (2007) argues that objective items in MC tests are flexible since they can assess "both global and detailed understanding of a text or focus..." (p. 18). Moreover, Simkin and Kuechler (2005) note that well-constructed MC items can overcome the critics on testing surface knowledge. On the contrary, they claim that quality MC questions can efficiently assess deep knowledge (Simkin & Kuechler, 2005).

It's also worth mentioning that MC tests are very reliable and objective since there is only one correct choice for each question (Coombe et al., 2007). Schuwirth and Van Der Vleuten (2004) argue that multiple choice tests are as much valid as open-ended exams since open-ended exams require intensive resource; and similarly MC tests are more reliable than open-ended exams since MC tests do not take much time to answer. Similarly, Bacon (2003) compared MC tests and short-answer (SA) questions and MC questions were found to be as reliable and valid as SA questions while being answered in a shorter time. The reliability is especially significant if a test is high-stake summative assessment and the score will affect the huge number of people (Bush, 2015). The reliability of a multiple choice test could be much more improved by the

number of the items (Bush, 2015). Oppenheim (2002) brings forth that although longer tests are more preferred than shorter tests in terms of validity, very long tests can tire and bother the students and prevent teachers to reach the real success of the students. However, as Burton (2006) states there is not a set criterion about how many items to include in the tests.

Familiarity of their format made MC tests even more appealing for students (Coombe et al., 2007). Most students have the ability to get better scores from an MC test rather than an essay type or open-ended test formats (Simkin & Kuechler, 2005). Moreover, students can get scores regardless of their writing abilities (Coombe et al., 2007) and they can guess the answer and get an extra point (Simkin & Kuechler, 2005), which could be argued to be an advantage of MC tests for students.

The studies conducted by Anderson and Krathwohl (as cited in Yonker, 2011) and Haladyna and Downing (1989) display that some Multiple-choice questions (MCQs) might test understanding and application of knowledge as well as the ability to understand the situations and solve problems demanding high-level thinking skills. This means that, MC tests do not assess just surface knowledge or learning as argued by Scouller (1998). Moreover, MC tests promote discriminative thinking skills since students are expected to choose among the plausible and competitive alternatives (Oppenheim, 2002).

Furthermore, in their experiments, Cantor et al. (2015) found out that MC testing “had the power to stabilize access to marginal knowledge” (p.193). The findings of their experiments proved that “marginal knowledge can easily be activated through multiple-choice test or re-exposure” (p.203). Hence, the student can remember many topics or subjects posed in the distractors even if that specific knowledge is not asked directly which can be argued as another benefit of MC tests.

Although multiple choice tests have been regarded as “necessary evils” in educational settings Little and Bjork (2015) disprove this baseless reputation advocating that properly created multiple choice tests with competitive alternatives seem to have a significant role in learning of not tested information on the alternatives. The above suggestions make significant contributions to our understanding of the reasons to prefer

MC tests in educational settings. However, different researchers argued some flaws or drawbacks of MC tests that cast shadow on their merits.

### **2.2.2. Criticism on multiple choice tests**

Despite their wide usage and many advantages, some studies argued that MC tests also have some drawbacks and weak points. Multiple-choice tests take criticism mostly about what they test. One of the most apparent disadvantages of MC tests is argued to be the fact that they are mostly used to test recognition (Walstad & Becker, 1994) although they can test higher-order thinking skills (Coombe et al., 2007). Likewise, Walsh and Seldomridge (as cited in DiBattista & Kurzawa, 2011) argue that the focus of MC tests is how well students can memorize the information rather than how much they can comprehend, practice or analyze the information. However, Gajjar et al. (2014) claim that MCQs assess the students' understanding and analyzing ability in addition to their knowledge. Also, Buckles and Siegfried (2006) clearly state that "although multiple choice questions cannot and should not be used to measure all levels of understanding, we believe that multiple-choice questions can test more than simple recognition and understanding..." (p.50). Moreover, MC tests are argued to fail to assess all parts of in-depth understanding such as synthesis and evaluation (Buckles & Siegfried, 2006).

As another significant downside of MC tests, it is argued that MC tests promote surface learning rather than deep learning (Scouller, 1998; Nicol, 2007). Students also think that multiple choice tests assess lower levels of thinking and ask knowledge based questions requiring the students just to remember that information (Scouller, 1998). Multiple choice questions only assess recall and they fail to assess higher-order thinking in addition to being written poorly Vahalia et al (1995). Moreover, Nicol (2007) argues that in an MC test, students are only supposed to recognize the most proper answer among the other options rather than to produce the answer. However, Gajjar et al. (2014) clearly state that "a good item can assess cognitive, affective as well as psychomotor domain (p.17). To overcome that handicap, Buckles and Siegfried (2006) suggest that the students might be asked to explain why they chose that option as a key answer and why they eliminated the other wrong options. Such kinds of explanations might enable MC tests to assess higher level of cognitive thinking.

Paxton (2000) criticizes the excessive emphasis on multiple choice tests and argues that multiple choice tests are unsuccessful in testing critical or communicative skills and because of the badly formed questions they also fail to measure problem-solving skills. In his analysis of an MC test exam taken by Economics department students, Paxton (2000) found that two-thirds of the questions were definition questions rather than application questions. His findings were parallel to Resnicks' (as cited in Paxton, 2000) in that multiple choice questions seem to test low level learning such as recalling rather than critical thinking ability. Similarly, Simkin and Kuechler (2005) argue that MC tests "fail to test students' ability to develop an argument" (p.2)

According to Shepard and Kirst (1991) MC tests have a tendency to assess lower level thinking skills and they are more vulnerable to guessing when compared to other types of assessment tools in which the students are supposed to construct their own sentences. According to Bush (2006) guessing might be encouraged by the overlapping questions in the tests. Teachers should be cautious in writing non-overlapping questions and providing the answer of another question incidentally. Moreover, Bush (2015) maintains that the sensitivity to guesswork puts the reputation of multiple choice tests on the line as an assessment tool. "A multiple choice item with five alternatives could be answered correctly by chance 20% of the time" (Olufemi & Oluseyi, 2012, p.239) unlike true-false items where the item could be answered correctly by chance half of the time. Schuwirth and Van Der Vleuten (2004) name this situation as "*cueing effect*" which means that the test taker recognizes the correct answer and find the key.

To get over these critics about guessing, most MC tests are currently scored with correction for guessing (deducting a percentage of a mark for each incorrect answer) or negative marking (deducting whole marks for incorrect answers) (Betts et al., 2009). Negative marking prevents the students from pure guessing since they will lose marks if they are unable to choose the correct answer (Betts, et al., 2009; Bush, 2015). Similarly, awarding the students for each right choice with +3 and punishing them for each incorrect choices with -1 is called "*pure-guess-neutral-scheme*" and guaranties that test takers will not gain or lose anything by just guessing (Bush, 2015, p.219). Moreover, Burton (2004) claims that in addition to deterring guessing, negative marking not also punishes misinformation but also enhances the reliability of the exam.

Getting artificial higher marks due to guessing is argued to be another main drawback of MC tests because a student can get a 20 or 25 % of the score per question with no knowledge (Bush, 2001). This affects both the overall results of the exams and the grades of the students individually. If the percentage of that test is significant for the overall end-year grades of the students that situation should not be neglected. According to Ahmadian et al., (2011) using “none of these” or “none of the above” as an option, minimizes the guessing factor, and thereby improves the test’s discrimination index and reliability. Bush (2001) suggests another solution to this problem introducing a new format of MC tests which is called as “*liberal*” (p.158).

In a “liberal” MC test students may choose more than one answer if they are not sure about the correct answer. Negative marking is used for each incorrect selection. In that way, students who have partial knowledge about the topic are rewarded against students who purely guess with no knowledge. As a result students are forced to think more carefully while answering the questions and that might be a better solution to guessing when compared to the traditional MC tests (Bush, 2001). This solution is parallel to Farthing’s (as cited in Bush, 2001) which argues that in order to reduce guessing by luck, MC tests may have more than one correct answer and students can get score combining those keys. Bush (2001) advocates that “liberal” MC tests with multiple correct answers assess knowledge more accurately than traditional MC tests with one correct answer; however, they should be used cautiously.

Due to the challenge and time cost in finding strong distractors and writing good questions most teachers might not favor MC tests (Coombe et al., 2007; Hansen & Dexter, 1997). Although they are easy to grade, objective items in MC tests force teachers (Brown & Hudson, 1998; Coombe et al., 2007). Instructors either spend too much time for writing good items and finding plausible distractors or spend less time and come up with a poor test not evaluating the target content (Hansen & Dexter, 1997). Both cases seem to bother teachers especially if they have no training in testing; thereby, these cases lead teachers to choose other assessment tools. However, Bacon (2003) advocates that, the time spent in preparing MC tests is not dependent on the number of the students; a huge number of students could be tested with a single MC tests and graded quickly. However, most teachers are supposed to prepare MC tests at

some point in their teaching practice regardless of their training, equipment or time (Goodrich, 1977).

Moreover, Gipps (as cited in Paxton, 2000) recalls the decreasing popularity of MC tests stating that MC tests are not being used as much as they used to be and that is because of their ill effects on the quality of teaching process and the curriculum. Paxton (2000) claims that, poorly written MC items carry disadvantages for the students who are not adept in verbal skills.

Another reason why MC tests are not welcomed might be that some students are testwise, which means they use some specific strategies for test format (Simkin & Kuechler, 2005). In MC tests, test takers' scores include two additional elements: partial knowledge combined with testwiseness and guessing among the options. Partial knowledge is different from guessing that is completely random (Burton, 2004). Students can eliminate some answers with their partial knowledge and get a higher score (Simkin & Kuechler, 2005). However, students can guess the answer even when they do not have a slightest idea about the topic that was asked (Biggs, 1999).

With respect to these two elements, partial knowledge and testwiseness, Rogers and Harley (1999) clearly states that:

It seems likely that these two additional components cannot be eliminated. People who construct multiple-choice items often find it difficult to construct items with a full set of plausible distractors or foils. The result is distractors that serve as test wise cues that test wise students can use to their advantages. It seems to follow then that if the number of the options is reduced, the influence of testwiseness would be reduced (p.236).

Despite the disadvantages and critics about the MC tests, it has been conclusively noted that “the simplicity and familiarity of traditional multiple choice tests will no doubt guarantee their continued long-term popularity for summative assessment” (Bush, 2015, p.229). Keeping in mind the widespread use of multiple-choice tests for assessment of students, item analysis of MC tests will provide teachers a rich pool of valid items to use along with their teaching practice.

### 2.3. Item Analysis

Items are regarded as the main components of most assessment types regardless of their type or length. An item “has always been the basic building block of a test” (Wainer, 1988, p. 2). “Test items are the foundation of tests and the backbone of most assessment instruments” (Coombe, et al., 2007, p.16). Items are important in the sense that they play significant roles in improving the reliability of tests (Burton, 2004). As Coombe, et al. (2007) advocate to make a test function well, all the other essential parts such as items, keys and the distractors need to work effectively.

As MC tests are increasingly preferred in assessing students’ learning, item analysis of MC tests has become a noteworthy research area in language teaching field (Ding & Beichner, 2009). Malau-Aduli and Zimitat (2012) stress the crucial role of item analysis stating that “evaluating the quality of any educational enterprise requires evaluation of the quality of the assessment within that system” (p. 919). In that sense, item analysis functions as a tool to control the quality of a test by eliciting numerical data at the item-level and the summary statistics of the whole test (Malau-Aduli & Zimitat, 2012). Moreover, Olufemi and Oluseyi (2012) suggest teachers and administrators to evaluate the exams before administering them to students.

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, et al., 2014; Olufemi & Oluseyi, 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 (Olufemi, & Oluseyi, 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; Olufemi & Oluseyi, 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).

“Item analysis is a simple yet valuable procedure performed after the examination providing information regarding the reliability and validity of an item\ test by calculating DIF I (Difficulty Index), DI (Discrimination Index), and DE (Distractor



Efficiency) and their relationship” (Gajjar, et al., 2014, p. 20) . Similarly, according to Cechova, et al., (2014) by item analysis the reliability and validity of a test are increased since it evaluates whether the items contribute to the objectives of the test or not. Gajjar et al. (2014) suggest that the capacity of an MC test rests on the quality of the tests which can be assessed by item analysis.

According to Brown (2004) preparing a test is a challenging job requiring science and art. Writing test items might be overwhelming even for the teachers who are adept in constructing test items. Poorly written items might include clues and increase the possibility of guessing incidentally (Burton, 2005). However, teachers or test developers ought to be cautious in that, writing too complicated items do not reduce guessing; it might be a myth rather than real (Burton, 2005). Downing (2005) states that reliability, and validity of assessment can be threatened by harmful effects of a poor quality item. A question may be interpreted differently by reader of the question despite all the carefulness and favorableness of the writer (Coombe, et al., 2007).

Item analysis also plays a significant role in improving teachers’ test construction skills (Olufemi & Oluseyi, 2012). According to Buckles and Siegfried (2006) if the items are written carefully, an MC test can assess higher-level cognitive processes, and they maintain that writing such items needs more skills when compared to writing items which are memory-based. However, Sproull (as cited in Oppenheim, 2002) claims that most questions developed by teachers or text banks require students only to remember information. Those poorly written MC items do not fit the learning goals such as solving problems, developing critical thinking, and applying and evaluating (Oppenheim, 2002).

Downing (2005) analyzed four MC tests and 219 items to examine the effects of flawed items on the tests in terms of reliability, item difficulty and item discrimination. In his study the items were defined either as ‘*standard*’ or ‘*flawed*’ if at least one or two of those item writing principles were violated. 100 of the items were found to be flawed. The findings of the study showed that although they were testing the same content the flawed items were 0-15% more difficult compared to the standard items. Moreover, 646 of the students passed the standard items while 575 students were able to pass the flawed items. 102 students (out of 749) passed the standard items but they failed the flawed items.

In that respect, flawed items penalized some students and reduced the validity and reliability of these exams. “It is likely that particular tests and with them their formats and scoring methods, have sometimes been judged as unreliable simply because of flawed items and procedures” (Burton, 2005, p. 66). Moreover, Ware and Vik (2009) argue that validity and reliability of a test depend on whether the items are written appropriately or not. Hence, due to the flawed items or violation of the guidelines, those students could be evaluated as unsuccessful and could fail the course.

Mehrens and Lehmann (as cited in Downing 2005) advocate that most of the tests prepared by instructors include many flawed items; however writing functional items might not be an easy job as Downing (2005) brings forth this difficulty stating that “test item writing may be as much art as science” (p.134). Moreover, it would be a “*misapprehension*” to think that all teachers have the ability to construct well-functioning and flawless items without taking any instruction (Burton, 2005).

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 serve to the testing aims.

### **2.3.1. Item facility**

Item facility (IF) is interchangeable with *item difficulty* or *facility value* and “refers to the proportion of the examinees who answered the question correctly, with lower values reflecting potentially more difficult questions” Osterlind (as cited in Malau-Aduli & Zimitat, 2012, p.921). Simply, it measures easiness of an item measure of although might be called as item difficulty (Ding & Beichner, 2009). Similarly, Downing (2005) defines item facility as the proportion of the examinees who answered the question correctly.

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; Olufemi & Oluseyi, 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, et al., 2014; Olufemi & Oluseyi, 2012). Item facility values might give information about the difficulty level of the topic that is tested (Álvarez, 2013). *P*-value can also function as a validity measure since the high values reflect the proportion of the students who have learned the content asked by the item (Malau-Aduli & Zimitat, 2012).

Item facility can be calculated by dividing the number of the students who answered the item correctly to the total number of the students answering the item. 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, 2004); .50 and .80 (Hamzah & Abdullah, 2011); .20 and .90 (Olufemi & Oluseyi, 2012); .30 and .80 (Coniam, 2009; Oppenheim, 2002); .30 and .70 (Brown, 2003); .50 and .90 (Haladyna & Downing, 1993).

According to Ebel (1967) if *p* value is equal to or above .40 that means this item needs no revision and should be kept for future use, if the value is between .29 and .20 this item should be revised, and if the value is below .20 these items need to be removed from the test. According to Doran (as cited in Ding & Beichner, 2009) items having difficulty level between .30 and .90 are satisfactory. Olufemi and Oluseyi (2012) maintain that very easy items which are above .90 are not worth testing and likewise, very difficult items below .20 can have confusing or ambiguous language or the content may be misstructured. Items having a difficulty level around .50 are ideal and they are the most reliable items (Brown, 2003; Ding and Beichner, 2009; Olufemi and Oluseyi, 2012). As opposed to these ideal ranges, Oppenheim (2002) advocates that the ideal difficulty level should be different for four-option items and five-option items. With four-option items the students can choose the key more easily when compared to the five-option items.

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 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.

To Wilson (as cited in Olufemi & Oluseyi, 2012) item difficulty is a fundamental factor in item analysis. Analyzing the difficulty of the items, whether the test takers have learned the topics asked in the items can be identified (Olufemi & Oluseyi, 2012). Bodner (1980) states that item facility indexes “do not indicate whether a question is good or bad, per se. They do, however, allow one to determine whether questions that one feels are trivial are truly trivial, or whether a question is difficult or truly impossible” (p. 189).

#### **2.3.1.1. Studies on item facility**

In their study conducted on 1198 items DiBattista and Kurzawa (2011) found a strong relationship between item facility and item discrimination. The most difficult items had very low discrimination index. Also, the discrimination index was higher for items having a difficulty index between .30 and .89. In other words, items having a difficulty index below .30 or above .90 were not likely to have a satisfactory discrimination index.

In another study carried by Gajjar et al. (2014) on 148 Medical School students, 50 items were analyzed and it was found that 24 items had good DIF Index (between .31 and .60) and they should have been stored for future use. 16 of the items were very difficult ( $\leq .30$ ) and 10 of the items were very easy ( $\geq .61$ ) and those 26 items should have been revised or discarded from the exam.

In their study with 40 items answered by 20 students, Hamzah and Abdullah (2011) found that 21 of the items were good in terms of difficulty (between .50 and .74); 9 of the items were easy items (between .25 and .49); 1 item was too easy (.11), and 2 of them were too difficult (.85 and above) and needed to be omitted from the test. Four of the easy questions were the earliest questions and they might have been prepared consciously to motivate the students (Hamzah & Abdullah, 2011).

After analyzing an MC test with 40 items taken by 800 students, Olufemi and Oluseyi (2012) found that the difficulty index of the most difficult item was .245 and only 76 students were able to answer it correctly. The difficulty index of the easiest item was .12 and it was answered correctly by 97% of the higher group and 85% of the lower group. Also, that item had a discrimination index of .12.

Ahmadian, et al., (2011) analyzed a 40-item receptive semantic prosody test and they found out that 8 of the items were difficult (.15 and .39) while 24 of them having a desirable level of difficulty (.40 and .70). However, 7 of the items were easy (.41 and .85), and only 1 of the items were very easy (.86 and .1), and needed to be replaced by another item for better future use of the test.

Analyzing a 50-item test taken by 63 students, Cechova, et al., (2014) noticed that 10 of the items were very easy (having a facility value .90 and above) and 2 of the items were too difficult (having a facility value .30 and below) for the students. Furthermore, they found out that those too difficult or too easy items did not help to discriminate the test takers.

Khodaday, et al., (2012) analyzed a 43-item achievement test based on schema theory (S-test) and they found out that 37 (68%) of the items were functioning well in terms of item facility (IF being between .25 and .75). These items also had good IF values which were .20 and higher. Moreover, only 2 of the items were too easy ( $\geq .90$ ) still they could be acceptable since they were the first items on the test.

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.

### **2.3.2. Item discrimination**

Item discrimination is also known as *discrimination index* (DI), *Index D* (Ebel, 1967) or *discriminative index* (Hamzah & Abdullah, 2011) and symbolized as “*d*”. DI value states how high achieving and low-achieving students answered the items within a test

(Malau-Aduli & Zimitat, 2012). Gajjar et al, (2014) define item discrimination as “the ability of an item to differentiate between students of higher and lower abilities” (p.18). It is also defined as difference in the percentages of correct responses to an item between the top quartile and the bottom quartile Oosterhof (as cited in Ding & Beichner, 2009).

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 is 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). Item discrimination is calculated as follows: the number of people in the upper group who answered the item correctly minus the number of people in the lower group who answered the item correctly, divided by the half of the number of the total of two groups (top and bottom groups). The number of the students in the top and the bottom groups are mostly equal and they represent one third of all test takers (Brown, J. D., 2003).

Item-discrimination indices are the numerical results obtained from the test data “that are used in assessing the effectiveness of the individual test items or questions” (Burton, 2001, p. 213). If an item has a good discrimination index it is supposed to be answered correctly more by the students from the top quartile rather than the students from the bottom quartile (Ding & Beichner, 2009). Item discrimination plays a significant role in the overall quality of the tests. If the items are able to discriminate highly between these two groups, they tend to yield more reliable results (Downing, 2005). Therefore, to prepare reliable tests the instructors should write items having a high discrimination index (Ebel, 1967).

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. For the top and bottom group the percentage have been used differently by researchers such as 25% (Gajjar, et al., 2014, Costin, 1972); 27% (Ebel, 1967; Goodrich, 1977; Jafarpur, 1999; Tarrant et al., 2009; Ware & Vik, 2009), 30% (Brown, 2004; Olufemi & Oluseyi, 2012, Kolstad et al., 1984 ) or 33%.

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).

In terms of evaluating the discrimination index of an item, different cut-off points have been suggested by researchers. For instance, according to Kolstad et al. (1984) the interpretation is as follows: .40 and above means excellent item, between .30 and .40 means good item, between .20 and .29 means average item and .10 and below means these items are unsatisfactory in terms of discrimination and they should be improved or omitted from the test. However, Jafarpur (1999) and Olufemi and Oluseyi (2012) define the items having .20 and higher indices as valuable. Doran (as cited in Ding & Beichner, 2009) and Coniam (2009) claim that .30 and above is an adequate index for discrimination. Furthermore, Ware and Vik (2009) claim that if an item has a discrimination index of  $<.15$  it means it does not have a discrimination power and the discrimination index of an item should be above .40 if it is said to discriminate excellently. Coombe, et al., (2007) argues that if the tests are reliable, the items should have an item discrimination value which is .30 and above. Moreover, Haladyna and Downing (1993) advocate that the discrimination index should be above .15.

Tian (2007) points out that “a good assessment method should be able to distinguish between deep learners and surface learners in a way so the former are rewarded while the latter are punished” (p. 387). Item discrimination has an important role in the reliability of a test. Ebel, (1967) points out that discriminatory power is a major determinant of the quality of an MC item. One of the presumptions in item discrimination is that “reliability of the test may be improved for future use by removing the items with low discrimination indices” (Burton, 2001, p. 213). If an item has low discrimination index the item should be investigated in terms of the clarity and wording of the question (Ding & Beichner, 2009). Ware and Vik (2009) declare that

“greater than or equal to 60% of items shall have moderate or better discrimination using set ranges” (p. 241).

One of the aims of the assessment is to identify the students who need help, who are progressing well, or who need more instruction (Coombe, et al., 2007). So, if an item cannot distinguish those students that may mean that the assessment is not working.

“For a MC item to have a good discriminatory power, examinees with higher test scores must select the keyed option more often than those with lower scores. For a distractor to be effective the opposite must be true- that is, examinees with higher test scores must select the distractor less often than those with lower scores” (DiBattista & Kurzawa, 2011, p. 3).

However, very easy items having weak distractor and not being able to discriminate between the high-achieving and the low-achieving groups can function as a warm-up if it is the first question on a test. (Coombe, et al., 2007; Gajjar et al., 2014)

A discriminating test is supposed to apparently distinguish between the students who have a strong knowledge about the asked item and the students who do not know the topic (Ding &Beichner, 2009). Discriminating items contribute to the reliability of the exam (Coombe, et al., 2007). Reliable tests are expected to differentiate between the students having different levels of proficiency on a domain (Oppenheim, 2002). Most of the time, the most discriminating items include the distractor about the misconception among the students, thereby help instructors to assess whether the students can still identify the key or not (Oppenheim, 2002). In that respect, the students are supposed to modify their knowledge to choose the key in this mélange of distractors.

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 is actually 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 an MC item fails to function effectively.

Bodner (1980) explains

In theory, the student who answers a given question correctly should have a tendency to perform better on the total examination than a student who answers the same question incorrectly. We therefore expect a positive correlation between the probability of a student getting a question right and the student's score on the exam. When the correlation coefficient for a correct answer is negative, something is drastically wrong with the question. (p.189)

### **2.3.1.1. Studies on item discrimination**

In their studies, DiBattista and Kurzawa (2011) analyzed 16 MC tests taken by from 109 to 547 students from different disciplines. After analyzing 1198 items they found out that, only 15% of the items were strong discriminators having a greater value than .40. More than 30% of the items had unsatisfactory discrimination values having a smaller value than .20 and also 4% of the items had negative discrimination values. Moreover, the finding of the study showed that the discrimination coefficients changed strikingly from one test to another. Additionally, it was seen that there was a strong relationship between discriminatory coefficients and the number of the functional distractors. In particular, as the number of functional distractors increased the discriminatory power of both four-option and five-option items improved. In other words, weak distractors had a dramatic effect on item discrimination and when the number of the weak distractors increased they could even cause the items to lose their

discriminatory powers at all (DiBattista & Kurzawa, 2011).

In the study by Gajjar et al. (2014) on Medical School students, it was found that only 15 items out of 50 had an excellent DI ( $\geq .25$ ); 9 of the items had a good DI (between .15 and .24), and the rest 26 items were very poor ( $<.15$ ) in terms of discriminating between the high-achieving and low-achieving students. The items were needed to be improved since they threatened the reliability of the test.

Olufemi and Oluseyi (2012) analyzed an MC test with 40 items taken by 800 students. They found that the discrimination index of 3 items was .0 and these items needed to be discarded from the test. Also, the discrimination index of the best discriminating item was .58. That item was answered correctly by 75% of the higher group and 17% of the lower group and was able to discriminate the groups well. It also had a moderate difficulty level like .46.

In another study, Hamzah and Abdullah (2011) analyzed an English language test taken by 20 students. They identified that out of 40 items, 9 of them were very good (.40 and above); 8 of them were good (between .30 and .40) and they should be kept; and 11 of them were unsatisfactory and poor items (.10 and below) in terms of discriminating the students in upper group and lower group. Their findings indicated that 10 of those 11 poor items had a very difficult and complex language and 1 of them was too easy. Those 11 items needed to be changed or discarded from the exam.

Cechova, et al. (2014) analyzed a 50 item test taken by 63 students and found out that, 26 of the items failed to discriminate among the students (having a discrimination value  $\leq .15$ ). Moreover, among these 26 items 2 of them had negative discrimination value ( $-.02$  and  $-.08$ ). 9 of the items were very good in terms of discrimination (having a discrimination value  $\geq .30$ ). They further suggested that those items should have been modified to increase the overall reliability of the test.

Tarrant et al. (2009) analyzed 514 items and their 2056 options in terms of discrimination indices and they found a strong relationship between the option discrimination and item discrimination. According to the findings of the study items having 3 functional distractors were the most discriminating items. As a result, it was

concluded that if an item had a distractor having a high discrimination value, that item was regarded as discriminating overall.

Costin (1972) analyzed the discriminatory power of three alternative items and four alternative items. In total he analyzed 220 items, and randomly half of the items were decreased to three alternatives. The results showed that the discrimination indices of three alternatives were higher than that of four alternatives, this study strongly concurs with Tversky (1964). The findings were disproving the idea that constructing more alternatives makes the test more discriminating and powerful.

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 discrimination value (-.09). They claimed that this item had the potential to create a negative washback effect for the students.

### **2.3.3. Distractor efficiency**

Distractors or disturbers are the options apart from the correct answer of a question in an MC test (Hamzah & Abdullah, 2011). Distractors are mostly included in the tests to trap the weaker students (Bodner, 1980) who did not grasp a specific concept and was not able to choose the key (Buckles & Siegfried, 2006). 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).

Distractors might be written according to common mistakes done by the students (Atalmış, 2014; Tarrant et al., 2009) or the misconceptions about the key (Haladyna & Downing, 1993; Oppenheim, 2002). Analyzing the distractors enables the researcher the opportunity to diagnose the general misconceptions among the students about a specific term or topic (Olufemi & Oluseyi, 2012; Buckles & Siegfried, 2006).

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. According to Ware and Vik (2009) "greater than or equal to 50% of all distractors shall be functioning at the 5% level" (p. 241). Nonfunctional distractors should be either replaced with a functioning one or be omitted from the test completely (Haladyna & Downing, 1989).

Writing functional or strong distractors has a crucial role in the overall quality of the tests because strong distractors increase the discriminatory power of the items (DiBattista & Kurzawa, 2011). Strong distractors are expected to resemble to the key, however choosing strong distractors might be a problematic issue for teachers. Rogers and Harley (1999) stress this problem stating that people constructing MC tests often have difficulty in constructing items with full of plausible distractors; and they mostly end up with distractors serving as cues to the test wise students. Tarrant, et al (2009) advocate that teachers spend little time on writing functioning distractors, and they spend much more time on forming the stems of the questions. However, functioning distractors are very significant parts in terms of the quality and reliability of a test (Haladyna & Downing, 1989).

Tarrant, et al. (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, et al., 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). Furthermore, teachers can write more items instead of writing more options, and this can able the teachers to cover the content more deeply (Tarrant et al., 2009).

Teachers or test developers might be aiming to diminish guessing while writing four or five alternatives. However, reducing the alternatives to three does not cause any significant statistical lost in the test in terms of discrimination (Costin, 1970; Ebel, 1969). According to Rich and Johanson (as cited in Atalmış, 2014) another way to write an item with fewer distractors is using “None of the above” (NOTA) as an alternative which works better than a weak distractor. NOTA can function as a competitive alternative and play a significant role in discriminating between the high-achieving and low-achieving students. Moreover, Farley (as cited in Tarrant et al., 2009) advocates that all questions do not have to include the same numbers of distractors; some questions might need more or less plausible distractors with respect to their content.

Great numbers of non-functional distractors can threaten the validity, reliability and accuracy of the test (Malau-Aduli & Zimitat, 2012). According to Malau-Aduli and Zimitat (2012) “ a distractor that fails to attract any examinees is dysfunctional, does not assist in the measuring of educational outcomes, adds nothing to the item or the test (psychometrically) and has negative impact upon learners” (p.927).

“The discriminatory power of a MC item depends heavily on the quality of its distractors. An effective distractor will look plausible to less knowledgeable students and lure them away from the keyed option; but it will not entice students who are well-informed about the topic under consideration” (DiBattista & Kurzawa, 2011, p.2).

Little and Bjork (2015) also emphasized the significance of plausible distractors stating that:

When multiple choice questions contain competitive incorrect alternatives, test-takers are led to retrieve previously studied information pertaining to all of the

alternatives in order to discriminate among them and select an answer, with such processing strengthening later access to information associated with both the correct and incorrect alternatives. (p.14)

In other terms, competitive alternatives can help the students learn the specific information that was asked and the competitive information placed in the alternatives (Little & Bjork, 2015). “The potency of a distractor must be combined with its factor of discrimination if its efficiency is to be determined; one quality without the other can be misleading” (Goodrich, 1977, p.70).

### **2.3.3.1. Studies on distractor efficiency**

In their study carried out on 3819 distractors, DiBattista and Kurzawa (2011) reported that only 54.8 % of the distractors were functional and more than one-third of the distractors were flawed because they were selected by less than 5% of the test takers. It was also found that, there was a strong relationship between item facility and strong distractors. According to the findings, more test takers chose the correct answer as the number of the strong distractors decreased. It was the same for both four-option items and five-option items on the tests.

Similarly, Tarrant et al. (2009) examined 514 items in terms of item analysis and they assessed the numbers of functioning and non-functioning distractors. The distractors that were chosen by less than 5% of the examinees were determined as non-functioning as Ware and Vik (2009), and Gajjar et al. (2014) suggest. They analyzed 2056 options; 1542 of them were distractors and 514 were the keys. The findings of their study revealed that 12.3% of the items had 0 functioning distractors. Also, 34.8% of the items had only 1 functioning distractors. Also, 39.1% of the items had 2 functioning distractors. The overall result of their study showed that only 52.2% of all distractors were functioning properly. The study concluded that the low number of the items having three plausible distractors indicate that teachers have difficulty in finding plausible distractors for four or five option items. Therefore it is suggested that, teachers might prefer three-option items instead of four or five.

Haladyna and Downing (1993) analyzed four multiple choice tests with 477 items. The tests had 2108 options in total, and 477 of them were correct answers while 1631 of

them were distractors. Their findings showed that over 38% of the distractors were non-functional and they were omitted from the tests since they were chosen less than 5 % of students. Moreover, they concluded that out of 200 items having five-options, none of them had four functional distractors. Hence, it is suggested that five-options do not serve to the aims of testing and teachers had better use their time and effort on forming two or three plausible distractors.

In their study Gajjar et al. (2014) analyzed 150 distractors in 50 MC items and reached that 133 of the distractors were functional and 17 of the distractors were non-functional distractors. Moreover, they inferred that items with functional distractors had higher discrimination indices. Designing plausible distractors might be argued to be a prerequisite for a quality test.

Costin (1972) analyzed a test having 100 items with four alternatives and randomly selected fifty of the items and reduced the number of alternatives to three instead of four. The test was administered to 1566 students as their final exam. As a result, three alternatives were found to be as valid and reliable as four alternatives and more efficient in terms of time and work load for teachers. Moreover, decreasing the number of alternatives did not hurt the test statistically; three alternatives were found to have the same or higher discrimination values as the four alternatives. Besides being less time consuming, three alternatives could also help to diminish the probability of guessing (Costin, 1972).

In their study conducted on 230 students, William and Ebel (as cited in Rogers & Harley, 1999) found out that the students answered the two or three-option items more quickly than the four-option items. Besides, two or three-option items had almost equal discrimination indexes.

Rogers and Harley (1999) used two forms of test with 40 multiple choice items. The first form had three option items and the second test had four option items, and the tests were taken by 158 students in total. They found out that when one of the ridiculous options was deleted, the tests were less susceptible in terms of testwiseness. The three option items tests were favored over the four option items tests.

Toksöz and Ertunç (2017) analyzed a 50-item multiple choice exam having 4 options and taken by 453 students in a language preparation class. They found that some of the distractors were insufficient and they were unable to attract any students from both groups (high achieving and low achieving). Moreover, they found that some distractors gathered more answers from high achieving students rather than low achieving students.

These three qualities namely item facility, item discrimination, and distractor efficiency are regarded as complementary; none of them is more significant or dominant than the others. Gajjar et al. (2014) stress the relationship between these three qualities stating that “more NFD in an item increase DIF (makes item easy) and reduces DE, conversely item with more functioning distractors decreases DIF (makes item difficult) and increases DE” (p.20). Olufemi and Oluseyi (2012) state that “each of this type of information serves a distinctive purpose which may be helpful to the conscientious teacher in improving both the teaching and testing procedures” (p.238).

Moreover, Oppenheim (2002) suggests that item discrimination is done to identify which students are lured by incorrect distractors and which students are not misled by the distractors. It is supposed that, distractors may attract the students having low level of achievement, whereas students having high level of achievements are not deviated by the distractors. Hence, these two analyses are said to be very closely related to each other. Furthermore, items having medium difficulty level has a good discrimination index and similarly the most discriminating items have medium difficulty level (Oppenheim, 2002). Therefore, the relationship between item facility and item discrimination seems to be ensured.

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, 1997; 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 MAKU which is a state university in Burdur, Turkey.



### **3. METHODOLOGY**

This chapter includes information on the participants, data collection process, and data collection tools and data analysis. Each section is aimed to present more detailed information about the design of the study.

#### **3.1. Research Design**

The research questions in the current study require both quantitative and qualitative data. Therefore, having a mixed method research design, the study employs both quantitative and qualitative data collection process. Integrating both types of data is needed to assist the researcher for data triangulation and transformation (Cresswell et al., 2004). Moreover, "...combining methods can open up fruitful new avenues for research in the social sciences" (Dörnyei, 2007, p.163).

One method may somehow overtake the other according to the aims or the research questions of the study. According to Dörnyei (2007), a study using mixed method has different sequence and dominance dimensions. In the light of his combinations, the present study includes the combinations of "QUAN → qual" (Dörnyei, 2007). That combination means that the quantitative research design comes first, and it is the dominant design because the analysis of the items of the multiple choice final exam is in the center of this research. Furthermore, quantitative design is followed by the qualitative research design to enrich the final findings adding depth to the statistical data and to validate the quantitative results with qualitative data namely interviews. Dörnyei (2007) advocates that the qualitative research design following the quantitative design "...puts flesh on the bones" (p. 45). Creswell et al. (as cited in Dörnyei, 2007) labels this combination as a '*sequential explanatory design*'.

#### **3.2. Participants**

The study was conducted with 210 non-compulsory preparatory school students studying at MAKU which is a state university in Burdur, Turkey. 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. These courses were taught by different lecturers and the students were not taking a specific course for Listening and Speaking skills. However, the students were also expected to listen and speak in other courses, especially Main Course and Reading & Writing courses. Also, audiovisual materials were used in the courses to aid the lecturers in developing students' listening and speaking skills.

### **3.3. Data Collection Process**

Quantitative data were collected through the final exams administered to the non-compulsory preparatory school students at MAKU in spring semester of 2014-2015 academic years. The necessary permission to use the exams in the study was obtained from the principal of the School of Foreign Languages at MAKU. After the appropriate institutional permissions were secured (See Appendix E), the quantitative data was collected right after the exam had been administered to the students. The students marked their answers on optical-scan answer sheets. Optical Mark Reader (OMR) was used to get the scores which were used by the researcher.

The qualitative data were collected through an interview prepared by the researcher and done by 21 of the test-takers who were non-compulsory preparatory school students at MAKU. The interview was conducted with the students who volunteered to aid the researcher. The interview was held one by one and face to face with the students at the researcher's office. The interviews lasted for 2 or 3 minutes approximately. The researcher prepared the interview both in Turkish and in English. Both versions were checked by two language experts in the field in terms of accuracy and clarity. However, the interview was done in Turkish to make the participants feel more comfortable in their mother tongue and express their ideas more clearly. The interview was recorded by the researcher after taking the consent from each participant. Later, the interviews were transcribed both in Turkish and English to be treated with content analysis.

### 3.4. Data Collection Tools

To collect quantitative data, the final exam session I (See Appendix A) and the final exam session II (See Appendix B) were used. The qualitative side of the study was strengthened through an interview with participants to get more in-depth information about their responses. With this aim, a semi-controlled interview was conducted with 10 percent (21) of the test takers. The interview was prepared both in Turkish (See Appendix C) and in English (See Appendix D).

#### 3.4.1. Final exams

To prevent the possibility of cheating from the other students four sets of exam papers (A, B, C, and D) with different sequence of questions were prepared. The final exam papers were professionally compiled as Coombe et al. (2007) noted; they were identified by a cover page with the name and content, date of the exam, duration, and version letter. The front page provided separate instructions for each test section such as cloze test, situation, and vocabulary. 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

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

As stated in Table 1, the final exam consists of five main parts: listening, grammar, vocabulary, reading (reading texts cloze test, conversation, situation, translation) and writing part. 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 were directly used by the researcher without modification.

The students were given 130 minutes to complete the exam. The questions were prepared and the exam was assembled by the instructors themselves. Also, the exam was reviewed by the instructors for the content, design and the classification of the items. Moreover, the final proofreading was also done by the instructors in terms of clearness or ambiguity.

The exams were held in two different sessions according to their times. The students were distributed to the sessions randomly. First session was held at 12:45 and the second session was held at 15:15. Both sessions had the same number of questions in each part such as 10 items in Listening part or 20 items in Grammar part. 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.

#### **3.4.2. Semi-structured interview**

The interviews gathered qualitative data and asked the students' ideas about the items and the distractors in the final exam of the non-compulsory preparatory school students at MAKU. In this study, a semi-structured interview was chosen because semi-structure interview helps the researcher direct the interview with respect to the research questions (Nunan, 1992). The interview was done in Turkish to get comparatively more illuminating information (See Appendix D). The interview included seven main questions and six follow-up questions. The interview questions were prepared with respect to the research questions of the study to see whether the quantitative and qualitative data confirm or conflict with each other. The questions were written in Yes/No questions format and open-ended format, regarding the participants' ideas about the test items, distractors and the exam in general. Only one characteristic was asked in each question.

### **3.5. Data Analysis**

#### **3.5.1. Quantitative data analysis**

While analyzing the quantitative data the students not selecting any of the options were eliminated to reach more accurate results. Since the analysis of item facility, item discrimination, and distractor efficiency are done according to the responses of the students, the papers of the students' not answering a question would be misleading for the 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 for item facility and item discrimination were derived from Brown (2004). Therefore, items having a difficulty level  $\geq .85$  were accepted as too easy items; items having a difficulty level  $< .15$  were accepted as too difficult; and items having a difficulty level between  $.15$  and  $.85$  were accepted as moderate difficult. With respect to item discrimination, items having a discrimination index  $\geq .50$  were accepted as highly discriminative. For distractor efficiency if a distractor is chosen by less than 5% of the test takers, it is regarded as non-functional distractor (NFD) (Gajjar et al. 2014; Ware & Vik, 2009; Downing & Haladyna, 1997). Some of the item characteristics such as readability, extremeness, social desirability, and content were neglected. In order to find out in which parts of the exam the students did well and in which parts the students did poorly the statistics program IBM SPSS Version 20 was used. Lastly, Paired-Samples T-test

was used in order to find out the relationship between the students' in-year grades point average and the students' final grades point average.

### **3.5.2. Qualitative data analysis**

The data collected through the interview was audio-recorded and transcribed by the researcher. According to Nunan (1992) tape recording has several strengths such as being naturalistic and objective. Also, it preserves actual language and it can be reanalyzed afterwards, and also participants' contributions can be recorded (Nunan, 1992). The transcriptions were treated with content analysis and interpreted by the researcher. However, the data and the interpretations in the study were limited to the participants' responses.

## 4. RESULTS

In this chapter, both quantitative and qualitative data results will be presented in line with the research questions. First, quantitative results will be given with the necessary tables and explanations. Secondly, qualitative results, namely the content analysis of the interview and some of the participants' extracts will be presented.

### 4.1. Quantitative results of the item analysis of the multiple choice final exams

#### 4.1.1. 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. Before presenting the tables it seems necessary to remind that the accepted cut-off points ( $\geq 85$  for too easy items,  $< 15$  for too difficult items, and  $.15$  and  $.85$  for moderate difficult items) were derived from Brown (2004).

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

| Item #   | <i>p</i> |
|----------|----------|
| Item # 1 | .85      |
| Item # 6 | .89      |
| Item # 8 | .94      |

According to Table 2, three items (3 %) in final exam session 1 are too easy ( $\geq 85$ ).

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

| Item #    | <i>p</i> | Item #    | <i>p</i> |
|-----------|----------|-----------|----------|
| Item # 10 | .10      | Item # 75 | .14      |
| Item # 13 | .12      | Item # 70 | .09      |
| Item # 28 | .12      | Item # 72 | .13      |
| Item # 32 | .15      | Item # 94 | .14      |

Table 3 shows that eight items (8 %) in final exam session 1 are too difficult ( $<15$ ).

Table 4. Item Facility (IF) indices of moderate difficult items in final exam session 1

| Item #    | <i>p</i> | Item #    | <i>p</i> | Item #    | <i>p</i> |
|-----------|----------|-----------|----------|-----------|----------|
| Item # 2  | .70      | Item # 35 | .80      | Item # 63 | .62      |
| Item # 3  | .45      | Item # 36 | .59      | Item # 64 | .54      |
| Item # 4  | .77      | Item # 37 | .66      | Item # 65 | .56      |
| Item # 5  | .58      | Item # 38 | .67      | Item # 66 | .22      |
| Item # 7  | .58      | Item # 39 | .29      | Item # 67 | .57      |
| Item # 9  | .46      | Item # 40 | .29      | Item # 68 | .39      |
| Item # 10 | .69      | Item # 41 | .53      | Item # 69 | .26      |
| Item # 11 | .46      | Item # 42 | .37      | Item # 71 | .40      |
| Item # 12 | .27      | Item # 43 | .25      | Item # 73 | .37      |
| Item # 14 | .43      | Item # 44 | .16      | Item # 74 | .25      |
| Item # 15 | .39      | Item # 45 | .17      | Item # 76 | .53      |
| Item # 16 | .37      | Item # 46 | .46      | Item # 77 | .49      |
| Item # 17 | .43      | Item # 47 | .37      | Item # 78 | .47      |
| Item # 18 | .43      | Item # 48 | .50      | Item # 79 | .35      |
| Item # 19 | .68      | Item # 49 | .23      | Item # 80 | .36      |
| Item # 20 | .53      | Item # 50 | .25      | Item # 81 | .46      |
| Item # 21 | .64      | Item # 51 | .42      | Item # 82 | .20      |
| Item # 22 | .48      | Item # 52 | .25      | Item # 83 | .60      |
| Item # 23 | .34      | Item # 53 | .24      | Item # 84 | .76      |
| Item # 24 | .53      | Item # 54 | .78      | Item # 85 | .31      |
| Item # 25 | .41      | Item # 55 | .78      | Item # 86 | .51      |
| Item # 26 | .33      | Item # 56 | .35      | Item # 87 | .29      |
| Item # 27 | .41      | Item # 57 | .25      | Item # 88 | .25      |
| Item # 29 | .56      | Item # 58 | .16      | Item # 89 | .31      |
| Item # 30 | .44      | Item # 59 | .41      | Item # 90 | .41      |
| Item # 31 | .17      | Item # 60 | .37      | Item # 91 | .48      |
| Item # 33 | .26      | Item # 61 | .57      | Item # 92 | .28      |
| Item # 34 | .19      | Item # 62 | .40      | Item # 93 | .39      |

Table 4 reveals that most of the items (57 %) in final exam session1 have moderate difficulty levels (between .15 and .85).

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

| Item #    | <i>p</i> |
|-----------|----------|
| Item # 8  | .92      |
| Item # 83 | .85      |

According to Table 5, two items (2 %) in final exam session 2 are too easy ( $\geq 85$ ).



Table 6. Item Facility (IF) indices of moderate difficult items in final exam session 2

| Item #    | <i>p</i> | Item #    | <i>p</i> | Item #    | <i>p</i> |
|-----------|----------|-----------|----------|-----------|----------|
| Item # 1  | .50      | Item # 34 | .26      | Item # 67 | .35      |
| Item # 2  | .22      | Item # 35 | .17      | Item # 68 | .40      |
| Item # 3  | .44      | Item # 36 | .38      | Item # 69 | .17      |
| Item # 4  | .32      | Item # 37 | .27      | Item # 70 | .40      |
| Item # 5  | .41      | Item # 38 | .20      | Item # 72 | .52      |
| Item # 7  | .61      | Item # 39 | .43      | Item # 73 | .22      |
| Item # 10 | .76      | Item # 40 | .16      | Item # 74 | .36      |
| Item # 12 | .24      | Item # 41 | .68      | Item # 75 | .29      |
| Item # 13 | .37      | Item # 42 | .52      | Item # 76 | .17      |
| Item # 14 | .21      | Item # 43 | .57      | Item # 77 | .37      |
| Item # 15 | .51      | Item # 44 | .21      | Item # 79 | .19      |
| Item # 16 | .35      | Item # 46 | .25      | Item # 80 | .32      |
| Item # 17 | .44      | Item # 47 | .54      | Item # 81 | .56      |
| Item # 18 | .50      | Item # 48 | .33      | Item # 82 | .45      |
| Item # 19 | .40      | Item # 50 | .22      | Item # 84 | .46      |
| Item # 20 | .57      | Item # 51 | .69      | Item # 85 | .44      |
| Item # 21 | .26      | Item # 52 | .34      | Item # 86 | .52      |
| Item # 23 | .30      | Item # 53 | .46      | Item # 87 | .74      |
| Item # 24 | .23      | Item # 54 | .36      | Item # 88 | .34      |
| Item # 25 | .54      | Item # 55 | .36      | Item # 90 | .28      |
| Item # 26 | .21      | Item # 56 | .23      | Item # 91 | .32      |
| Item # 27 | .24      | Item # 58 | .50      | Item # 92 | .56      |
| Item # 28 | .34      | Item # 59 | .30      | Item # 93 | .31      |
| Item # 29 | .30      | Item # 60 | .30      | Item # 94 | .57      |
| Item # 30 | .21      | Item # 61 | .19      | Item # 95 | .47      |
| Item # 31 | .25      | Item # 63 | .23      |           |          |
| Item # 32 | .49      | Item # 65 | .36      |           |          |
| Item # 33 | .19      | Item # 66 | .20      |           |          |

Table 6 demonstrates that most of the items (86 %) in final exam session 2 have moderate difficulty levels (between .15 and .85).

Table 7. Item Facility (IF) indices of too difficult items in final exam session 2

| Item #    | <i>p</i> | Item #    | <i>p</i> | Item #    | <i>p</i> |
|-----------|----------|-----------|----------|-----------|----------|
| Item # 6  | .15      | Item # 49 | .12      | Item # 71 | .14      |
| Item # 9  | .03      | Item # 57 | .05      | Item # 78 | .10      |
| Item # 11 | .09      | Item # 62 | .15      | Item # 89 | .14      |
| Item # 45 | .14      | Item # 64 | .10      |           |          |

According to Table 7, 11 items (11.5 %) in final exam session 2 are too difficult (<.15).

Table 8. Overall Item Facility (IF) results of final exams

|           | <i>Too easy</i> | Moderate | Too difficult |
|-----------|-----------------|----------|---------------|
| Session 1 | 3               | 84       | 8             |
| Session 2 | 2               | 82       | 11            |

Table 8 demonstrates that in session 1 there were 3 too easy items, 8 too difficult items and 84 moderate difficult items. In session 2, there were 2 too easy items, 11 too difficult items, and 82 moderate difficult items. Hence, it might be argued that the final exams had mix of difficulty levels and most items had acceptable difficulty levels.

#### 4.1.2. Item discrimination indices of the items on the final test exams

In this part, the item discrimination indices (DI) of the items in final exam session 1 and final exam session 2 will be presented in tables and analyzed. It seems necessary to remind that the cut-off points (below 0 for negative discrimination, below .30 for zero or low discrimination, and .30 and above for high discrimination) were derived from Brown (2004).

Table 9. Items with negative discrimination indices in final exam session 1

| Item #    | <i>DI</i> | Item #    | <i>DI</i> |
|-----------|-----------|-----------|-----------|
| Item # 4  | -0.01     | Item # 55 | -0.02     |
| Item # 6  | -0.01     | Item # 60 | -0.06     |
| Item # 75 | -0.02     | Item # 94 | -0.04     |

Table 9 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. To illustrate, the distribution of responses for item # 60 in final exam session 1 is shown in Table 9 below.

Table 10. Distribution of responses for item # 60 in final exam session 1

| Item # 60                | <i># Correct</i> | <i>#Incorrect</i> |
|--------------------------|------------------|-------------------|
| High Ability Ss (Top 36) | 10               | 26                |
| Low Ability Ss (Top 36)  | 15               | 21                |

Table 10 demonstrates that item # 60 in Cloze Test part of the exam gathered more correct answers from low ability students rather than high ability students. The item

needs to be modified or revised to prevent negative washback effect for high ability students.

Table 11. Items with moderate discrimination indices in final exam session 1

| Item #    | <i>DI</i> | Item #    | <i>DI</i> |
|-----------|-----------|-----------|-----------|
| Item # 15 | 0.38      | Item # 61 | 0.33      |
| Item # 36 | 0.38      | Item # 91 | 0.37      |
| Item # 48 | 0.37      | Item # 93 | 0.30      |

Table 11 demonstrates that 6 items (6.31%) in final exam session 1 had moderate discrimination indices. 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.

Table 12. Items with zero or low discrimination indexes in final exam session 1

| Item #    | <i>DI</i> | Item #    | <i>DI</i> | Item #    | <i>DI</i> |
|-----------|-----------|-----------|-----------|-----------|-----------|
| Item # 1  | 0.08      | Item # 32 | 0.16      | Item # 65 | 0.20      |
| Item # 2  | 0.09      | Item # 33 | 0.08      | Item # 66 | 0.02      |
| Item # 3  | 0.04      | Item # 34 | 0.15      | Item # 67 | 0.26      |
| Item # 5  | 0.08      | Item # 35 | 0.20      | Item # 68 | 0.15      |
| Item # 7  | 0.02      | Item # 37 | 0.11      | Item # 69 | 0.05      |
| Item # 8  | 0.04      | Item # 38 | 0.29      | Item # 70 | 0.01      |
| Item # 9  | 0.01      | Item # 39 | 0.22      | Item # 71 | 0.23      |
| Item # 10 | 0.09      | Item # 40 | 0.04      | Item # 72 | 0.01      |
| Item # 11 | 0.09      | Item # 41 | 0.25      | Item # 73 | 0.22      |
| Item # 12 | 0.08      | Item # 42 | 0.22      | Item # 74 | 0.04      |
| Item # 13 | 0.08      | Item # 43 | 0.06      | Item # 76 | 0.18      |
| Item # 14 | 0.22      | Item # 44 | 0.15      | Item # 77 | 0.26      |
| Item # 16 | 0.29      | Item # 45 | 0.02      | Item # 78 | 0.19      |
| Item # 17 | 0.12      | Item # 46 | 0.27      | Item # 79 | 0.22      |
| Item # 18 | 0.01      | Item # 47 | 0.13      | Item # 80 | 0.23      |
| Item # 19 | 0.16      | Item # 49 | 0.12      | Item # 81 | 0.25      |
| Item # 20 | 0.19      | Item # 50 | 0.20      | Item # 82 | 0.05      |
| Item # 21 | 0.23      | Item # 51 | 0.12      | Item # 83 | 0.18      |
| Item # 22 | 0.25      | Item # 52 | 0.05      | Item # 84 | 0.25      |
| Item # 23 | 0.15      | Item # 53 | 0.01      | Item # 85 | 0.25      |
| Item # 24 | 0.22      | Item # 54 | 0.15      | Item # 86 | 0.16      |
| Item # 25 | 0.13      | Item # 56 | 0.29      | Item # 87 | 0.12      |
| Item # 26 | 0.26      | Item # 57 | 0.18      | Item # 88 | 0.23      |
| Item # 27 | 0.25      | Item # 58 | 0.13      | Item # 89 | 0.11      |
| Item # 28 | 0.04      | Item # 59 | 0.06      | Item # 90 | 0.16      |
| Item # 29 | 0.18      | Item # 62 | 0.20      | Item # 92 | 0.05      |
| Item # 30 | 0.19      | Item # 63 | 0.25      | Item # 95 |           |
| Item # 31 | 0.15      | Item # 64 | 0.27      |           |           |

Table 12 shows that 83 items (87.36 %) had discrimination values which are zero or very low (<30) thereby these items were unable to meet the requirements.

Table 13. Items with negative discrimination indices in final exam session 2

| Item #    | <i>DI</i> | Item #    | <i>DI</i> |
|-----------|-----------|-----------|-----------|
| Item # 6  | -0.07     | Item # 57 | -0.04     |
| Item # 9  | -0.01     | Item # 60 | -0.02     |
| Item # 12 | -0.02     | Item # 62 | -0.04     |
| Item # 17 | -0.13     |           |           |

According to Table 13, 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 in Table 13, Table 14, and Table 15 below.

Table 14. Distribution of responses for item # 6 in final exam session 2

| Item #                   | <i>#Correct</i> | <i>#Incorrect</i> |
|--------------------------|-----------------|-------------------|
| High Ability Ss (Top 36) | 5               | 29                |
| Low Ability Ss (Top 36)  | 10              | 14                |

Table 14 shows that most high ability students failed to answer item # 6 correctly in final exam session 2. Low ability students were more successful for that item contrary to the expectations.

Table 15. Distribution of responses for item # 17 in final exam session 2

| Item #                   | <i># Correct</i> | <i>#Incorrect</i> |
|--------------------------|------------------|-------------------|
| High Ability Ss (Top 36) | 9                | 25                |
| Low Ability Ss (Top 36)  | 18               | 16                |

As seen in Table 15, more students from low ability group rather than high ability group were able to answer item # 17 correctly in final exam session 2. The item needs to be improved to prevent the possible negative washback effect for students.

Table 16. Distribution of responses for item # 57 in final exam session 2

| Item #                   | # Correct | #Incorrect |
|--------------------------|-----------|------------|
| High Ability Ss (Top 36) | 0         | 34         |
| Low Ability Ss (Top 36)  | 3         | 31         |

Table 16 shows that none students from high ability group was able to answer item # 57 in final exam session 2 correctly. However, 3 students from low ability group could answer the item correctly. These items having negative item discrimination indices are probable 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 17. Items with zero or low discrimination indexes in final exam session 2

| Item #    | DI   | Item #    | DI   | Item #    | DI   |
|-----------|------|-----------|------|-----------|------|
| Item # 1  | 0.22 | Item # 32 | 0.27 | Item # 64 | 0.10 |
| Item # 2  | 0.01 | Item # 33 | 0.16 | Item # 65 | 0.07 |
| Item # 3  | 0.14 | Item # 34 | 0.02 | Item # 66 | 0.14 |
| Item # 4  | 0.16 | Item # 35 | 0.14 | Item # 67 | 0.10 |
| Item # 5  | 0.05 | Item # 36 | 0.16 | Item # 68 | 0.22 |
| Item # 7  | 0.22 | Item # 37 | 0.26 | Item # 69 | 0.07 |
| Item # 8  | 0.08 | Item # 38 | 0.17 | Item # 70 | 0.22 |
| Item # 10 | 0.13 | Item # 39 | 0.25 | Item # 71 | 0.01 |
| Item # 11 | 0.01 | Item # 40 | 0.14 | Item # 73 | 0.11 |
| Item # 13 | 0.02 | Item # 41 | 0.25 | Item # 74 | 0.10 |
| Item # 14 | 0.26 | Item # 42 | 0.23 | Item # 75 | 0.04 |
| Item # 15 | 0.27 | Item # 44 | 0.01 | Item # 76 | 0.01 |
| Item # 16 | 0.25 | Item # 45 | 0.04 | Item # 77 | 0.09 |
| Item # 18 | 0.23 | Item # 46 | 0.23 | Item # 78 | 0.01 |
| Item # 19 | 0.23 | Item # 48 | 0.27 | Item # 79 | 0.02 |
| Item # 20 | 0.26 | Item # 49 | 0.07 | Item # 80 | 0.11 |
| Item # 21 | 0.00 | Item # 50 | 0.16 | Item # 81 | 0.16 |
| Item # 22 | 0.16 | Item # 51 | 0.14 | Item # 82 | 0.07 |
| Item # 23 | 0.13 | Item # 52 | 0.05 | Item # 83 | 0.16 |
| Item # 24 | 0.14 | Item # 53 | 0.10 | Item # 84 | 0.14 |
| Item # 25 | 0.23 | Item # 54 | 0.11 | Item # 85 | 0.29 |
| Item # 26 | 0.10 | Item # 55 | 0.14 | Item # 87 | 0.25 |
| Item # 27 | 0.13 | Item # 56 | 0.04 | Item # 88 | 0.16 |
| Item # 28 | 0.13 | Item # 58 | 0.19 | Item # 89 | 0.01 |
| Item # 29 | 0.14 | Item # 59 | 0.13 | Item # 90 | 0.10 |
| Item # 30 | 0.13 | Item # 61 | 0.14 | Item # 91 | 0.14 |
| Item # 31 | 0.08 | Item # 63 | 0.10 |           |      |

Table 17 shows that 80 items (84.21 %) had discrimination values which are zero or very low (<0.30) thereby these items were unable to meet the requirements.

Table 18. Items with moderate discrimination indices in final exam session 2

| Item #    | <i>DI</i> | Item #    | <i>DI</i> |
|-----------|-----------|-----------|-----------|
| Item # 43 | 0.30      | Item # 92 | 0.33      |
| Item # 47 | 0.35      | Item # 93 | 0.33      |
| Item # 72 | 0.32      | Item # 94 | 0.32      |
| Item # 86 | 0.32      | Item # 95 | 0.30      |

Table 18 demonstrates that 8 items (8.42 %) in final exam session 2 had acceptable discrimination indexes (.30 and above) (Coombe, et al., 2007). However, Brown (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, 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.

Table 19. Overall Discrimination Index (DI) results of final exams

|           | <i>Negative</i> | <i>Zero or low</i> | Moderate |
|-----------|-----------------|--------------------|----------|
| Session 1 | 6               | 83                 | 6        |
| Session 2 | 7               | 80                 | 8        |

Table 19 shows that none of the items in final exam session and final exam session 2 had high discrimination index. The number of the items having moderate difficulty levels was very small. Hence, it can be said that the exams were suffering in terms of discriminating between high ability and low ability students.

#### **4.1.3. Distractor efficiency of the options of the items on the final test exams**

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. Before, presenting the tables it seems necessary to remind some important details that need to be kept in mind. First, 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). Second, 108 students took final exam session 1, and 102 students took final exam sessin 2. Third, the Listening

parts of both sessions had 3 options while the other parts in both sessions such as Vocabulary or Grammar had 5 options.

Table 20. Items with NFDs in Listening part of final exam session 1

| Item # | A  | B         | C          | D | E |
|--------|----|-----------|------------|---|---|
| # 1    | 11 | 5         | <b>92</b>  | - | - |
| # 6    | 3  | <b>97</b> | 8          | - | - |
| # 8    | 4  | 2         | <b>102</b> | - | - |

Note. Bold options are the correct answers

Table 20 shows that 3 items (30%) in Listening part of final exam session 1 had NFD distractors.

Table 21. Items with NFDs in Grammar part of final exam session 1

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

Note. Bold options are the correct answers

Table 21 demonstrates that 8 items (40%) in Grammar part of final exam session 1 had NFD distractors.

Table 22. Items with NFDs in Vocabulary part of final exam session 1

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

Note. Bold options are the correct answers

Table 22 shows that 5 items (25%) in Vocabulary part of final exam session 1 had NFD distractors.

Table 23. Items with NFDs in Cloze Test part of final exam session 1

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

Note. Bold options are the correct answers

Table 23 shows that 5 items (50%) in Cloze test part of final exam session 1 had NFD distractors.

Table 24. Items with NFDs in Translation part of final exam session 1

| Item # | A  | B  | C         | D         | E  |
|--------|----|----|-----------|-----------|----|
| # 61   | 16 | 13 | 4         | <b>62</b> | 13 |
| # 65   | 5  | 10 | <b>61</b> | 10        | 22 |

Note. Bold options are the correct answers

Table 24 demonstrates that 2 items (33.3%) in Translation part of final exam session 1 had NFD distractors.

Table 25. Items with NFDs in Reading part of final exam session 1

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

Note. Bold options are the correct answers



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

All the distractors in Dialogue and 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 final exam session 1. Hence, the tables of these parts are not presented here.

Overall, 10% of the distractors in 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 final exam session 1. 90% of the distractors were found to function properly. Hence, it can be argued that the exam had strong distractors in general. However, to improve the validity and reliability of the exam the non-functional distractors should be modified.

Table 26. Items with NFDs in Listening part of final exam session 2

| Item # | A         | B        | C | D | E |
|--------|-----------|----------|---|---|---|
| # 8    | <b>94</b> | 4        | 4 | - | - |
| # 9    | 89        | <b>4</b> | 9 | - | - |

Note. Bold options are the correct answers

Table 26 shows that 2 items (20%) in Listening part of final exam session 2 had NFD distractors.

Table 27. Items with NFDs in Grammar part of final exam session 2

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

Note. Bold options are the correct answers

Table 27 demonstrates that 4 items (20%) in Grammar part of final exam session 2 had NFD distractors.

Table 28. Items with NFDs in Vocabulary part of final exam session 2

| Item # | A         | B  | C | D         | E         |
|--------|-----------|----|---|-----------|-----------|
| # 41   | 17        | 6  | 2 | <b>70</b> | 7         |
| # 47   | <b>56</b> | 33 | 8 | 3         | 2         |
| # 49   | 11        | 36 | 4 | 38        | <b>13</b> |

Note. Bold options are the correct answers

Table 28 shows that 3 items (15%) in Vocabulary part of final exam session 2 had NFD distractors.

Table 29. Items with NFDs in Cloze Test part of final exam session 2

| Item # | A         | B  | C | D  | E  |
|--------|-----------|----|---|----|----|
| # 51   | <b>71</b> | 12 | 4 | 4  | 11 |
| # 53   | <b>47</b> | 4  | 7 | 32 | 12 |

Note. Bold options are the correct answers

Table 29 demonstrates that 2 items (20%) in Cloze test part of final exam session 2 had NFD distractors.

Table 30. Items with NFDs in Reading part of final exam session 2

| Item # | A         | B | C  | D  | E |
|--------|-----------|---|----|----|---|
| # 65   | <b>37</b> | 8 | 32 | 21 | 4 |

Note. Bold options are the correct answers

Table 30 shows that only 1 item (5%) in Reading part of final exam session 2 had NFD distractors.

Table 31. Items with NFDs in Translation part of final exam session 2

| Item # | A  | B         | C         | D | E  |
|--------|----|-----------|-----------|---|----|
| # 82   | 38 | <b>46</b> | 3         | 4 | 11 |
| # 83   | 4  | 3         | <b>87</b> | 5 | 3  |

Note. Bold options are the correct answers

Table 31 demonstrates that 2 items (33.3%) in Translation part of final exam session 2 had NFD distractors.

Table 32. Items with NFDs in Dialogue part of final exam session 2

| Item # | A  | B         | C  | D         | E |
|--------|----|-----------|----|-----------|---|
| # 87   | 13 | <b>76</b> | 5  | 6         | 2 |
| # 91   | 30 | 17        | 20 | <b>33</b> | 2 |

Note. Bold options are the correct answers

Table 32 shows that 2 items (22.2%) in Dialogue part of final exam session 2 had NFD distractors.

All the distractors in 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 of 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 final exam session 2. Almost 93% of the distractors were found to function properly. Therefore, it may be argued that the exam had strong distractors in general. However, to improve the validity and reliability of the exam the non-functional distractors should be revised.

#### 4.1.4. The success of the students according to the parts of the exam

In this part, in which parts of the exam the students did well and in which parts the students did poorly will be presented in tables and analyzed.

Table 33. The success of the students according to the parts of the final exam session 1

| Part of the exam | Total Point* | Success |
|------------------|--------------|---------|
| Listening        | 712          | 66%     |
| Grammar          | 929          | 43%     |
| Vocabulary       | 816          | 38%     |
| Cloze Tests      | 390          | 36%     |
| Reading          | 846          | 39%     |
| Translation      | 321          | 49.5%   |
| Dialogue         | 139          | 32%     |
| Situation        | 193          | 36%     |

\*108 students took final exam session 1

Table 33 demonstrates that students who participated in final exam session 1 were most successful (66%) in Listening part of the final exam and they did worst (32%) in Dialogue part of the exam.

Table 34. The success of the students according to the parts of the final exam session 2

| Part of the exam | Total Point* | Success |
|------------------|--------------|---------|
| Listening        | 450          | 44%     |
| Grammar          | 693          | 34%     |
| Vocabulary       | 662          | 32%     |
| Cloze Tests      | 370          | 36%     |
| Reading          | 543          | 27%     |
| Translation      | 337          | 55%     |
| Dialogue         | 202          | 49.5%   |
| Situation        | 183          | 36%     |

\*102 students took final exam session 2

Table 34 shows that students who participated in final exam session 2 were most successful (55%) in Translation part of the final exam and they did worst (32%) in Vocabulary part of the exam.

#### 4.1.5. Non-compulsory preparatory school students' in-year grades point average and the students' final grades point average?

In this part, the relationship between the non-compulsory preparatory school students' in-year grades point average and the students' final grades point average will be presented and analyzed.

Table 35. T-test results for students' in-year grades and final grades point average

| Grades  | N   | $\bar{x}$ | SD    | <i>df</i> | t      | p    |
|---------|-----|-----------|-------|-----------|--------|------|
| In-year | 266 | 52.06     | 16.01 | 265       | 27,134 | .000 |
| Final   | 266 | 38.87     | 15.04 |           |        |      |

Table 35 shows that there is a significant difference ( $p = .000$ ) in participants' final grades according to in-year grade averages ( $p \leq .05$ ). That means the students' in-year grades were higher than the students' final grades. The exams administered during the year might be easier than the final exam. The students might be mentally loaded at the

end of the year and they might have done worse in final exam. Since final exams have more weight in students' overall score, the students might be more anxious during the exam and that might have decreased their success.

#### **4.2. Non-compulsory preparatory school students' perceptions about the difficulty level of the items, the discriminatory power of the exam, the efficiency of the distractors and the exam in general**

In this part, interview results which were subject to content analysis will be presented. The results will be given along with the order of the research questions.

##### **4.2.1. Participants' remarks about the difficulty level of the items in the final exam**

To find out what the students think about the difficulty level of the items on the tests they were asked the following question: How was the difficulty level of the multiple choice questions of the final exam? Their answers show the students have different remarks about the difficulty of the exam. For instance, 28% of the students think that the exam was very difficult:

P4: It was really difficult, I mean I was torn between two options all the time; and some questions were too long. Yes, it was very difficult.

P14: I think the questions were too difficult according to our levels. There were question from every nook and cranny, and they forced us.

Some participants (28 %) think that the exam was difficult but it was not too difficult:

P6: Since we are non-compulsory preparatory school students, the exam was difficult for us. It was not easy.

P3: It was difficult when compared to the other exams that we have taken so far during the year.

Still some participants (28%) think that the exam had moderate difficulty:

P10: It was neither difficult nor easy. The exam was very suitable for our levels.

P11: It had a moderate difficulty.

On the other hand, some participants (14%) think that the exam was easy:

P13: I think the exam was not difficult for a student who studied before. It was very good, I mean it was easy.

P18: The exam was easy but I hadn't studied for it.

#### **4.2.2. Participants' remarks about the discriminatory power of the items in the final exam**

To understand the discriminatory power of the items on the tests the students are asked: What can you tell about the discriminating power of the exam? Could the exam discriminate well between the good and poor or hardworking and lazy students? Based on the responses elicited from the participants it has been found that most of the students (90.4%) think that the exam was very discriminative for the students:

P2: The exam could discriminate well between the students who studied and who did not, because the questions had different difficult levels.

P5: Since it was challenging for me although I prepared well for the exam, it must be discriminative enough. I mean the ones who did not study for the exam could not do it.

P17: Yes, it had a discriminatory power. The ones who studied hard could do it, we did not study hard and we sweated.

P21: There were many discriminative questions. It was designed to eliminate the students who did not know the topic. I think only the ones who study English for two or three hours a day could score 70 points or over.

However, a few students (10 %) stated that the exam was not discriminative:

P4: It could not discriminate well. It was too comprehensive. All our friends who studied and prepared well for the exam sweated too like us.

P6: It was not discriminative. We would get low grades even if we studied hard, and we could get low grades again although we studied hard.

#### **4.2.3. Participants' remarks about the efficiency of the distractors in the final exam**

To investigate how effective were the distractors on the tests the students were asked the following questions: How was the efficiency of the distractors? Were you ever torn between two options (choices)? Why? With reference to participants' quotations, it has been found out that most of the participants (76 %) think that the distractors were very efficient and they were often torn among the options:

P1: The questions in translation parts had very good distractors. The options were very close and similar to each other. Yes, I was torn between two options.

P2: The distractors were too effective. They were all similar to each other especially in translation part. I had difficulty to choose the answer because the options were very similar.

P6: Some options were very similar. We could have not answered them only because we did not see the difference. Yes, I was really torn between the options.

P13: I was usually torn between two options. I could eliminate the other three options though. About being torn between the options, if you studied hard, you could do it. I tried to choose the best option.

P14: The options were designed to distract the students. At least two options looked almost the same. Yes, I was torn between the options. The options were very similar; our teachers were very vigilant about it.

In spite of the fact that most students think that the distractors were efficient. Some participants (10%) think that it depended on the part of the exam:

P5: Sometimes I was torn between two options but still some of the keys were very obvious. For example, in vocabulary and listening parts, the correct answers were easily noticed; however, in grammar parts the distractors were very effective.

P21: In grammar part, the distractors were very strong; however, in vocabulary part the answer was very obvious. They were very easy, and the options were different from the key. Yes, I was torn between the options especially in reading and grammar parts. For reading part I could not understand the paragraphs and for grammar part, I confused some of the grammatical rules.

On the other hand some participants (14%) stated that the distractors failed and they could not fool them:

P11: No, not at all. The keys were very obvious especially for hardworking students. They were easy for me.

P19: Actually, they were good; however I was not torn between two choices. They could not fool me.

#### **4.2.4. Participants' remarks about their feelings and motivations when they were torn between two options**

To investigate the participants' feelings and motivations when they were torn between two options the students were asked the following questions: How did you feel when you were torn between two options? Did the questions that you were torn between two

options decrease your motivation while answering the other questions in the exam? Their answers show that most students felt very bad and helpless when they were torn between the options:

P1: I felt terrible. I thought to give up and go home.

P9: I was about to cry. I wanted to cry.

P15: I was very excited. I thought that I would not do it.

Few students mentioned in their remarks that they did not feel very bad and handled the problem easily:

P6: I read all the options and chose one which is best for me.

P8: I skipped that question and answered the next one.

P14: I had a 50% chance, so I turned the wheel, you know.

When it comes to motivation, participants' remarks revealed that most students' motivation decreased when they were torn between two options:

P7: They decreased my motivation because I thought that I could not finish the exam in time.

P9: Yes, it did. I did not want to have the makeup exam and I was afraid that I would.

P21: Yes, it decreased my motivation because they were always in my mind during the whole exam. And there were many of them.

Few participants remarked that their motivation was not affected when they were torn between two options.

P12: No, it did not. I was relaxed.

P4: Whenever I was torn between two options, I skipped the question. I let myself to turn back again to that question. I answered the other questions.

#### **4.2.5. Participants' remarks about the most difficult and easiest part of the final exam**

To find out in which parts of the exams the students did poorly and in which parts of the exams the students did well, the students were asked: In the multiple choice part of the exam, there were four main parts: listening skills, vocabulary, grammar and reading comprehension.

Which part was the most difficult for you? Why? Which part was the easiest for you? Why? Participants' remarks show that the students had difficulty mostly in reading comprehension part of the final exam:



P1: I had difficulty in paragraphs mostly. I had to read many times to understand. They were too long or I did not the vocabularies in them. I had difficulty in understanding the questions.

P13: The paragraphs were too long and I was bored. And also, I had difficulty in the questions finishing with the word “say”.

P21: The reading part was the most difficult part for me. The paragraphs were very long. The options were confusing. The paragraphs were too difficult when compared to the paragraphs that we came across during the lessons. 4 or 5 questions were related to the same paragraph and I think that is too much.

Moreover, most of the students had difficulty in vocabulary part of the final exam:

P6: I had difficulty mostly in vocabulary parts. There were too many unknown words for us.

P11: The vocabulary part was the most difficult one for me. I am not good at learning or memorizing the new words.

Furthermore, 2 students mentioned in their remarks that they had difficulty mostly in listening parts and 2 students remarked that the grammar part was the most difficult part for them.

P9: Listening part was very difficult for me. The speakers were at an airport and there were too much noise and I could not hear them clearly.

P15: I had difficulty mostly in listening part of the exam. I think they were related to the videos.

P16: Grammar part was very difficult for me because I did not study hard.

P20: Grammar part was very difficult for me because there were unknown words.

When it comes to the easiest part of the final exam, most participants mentioned that the easiest part for them was translation part in the final exam.

P16: Translations were very easy because they had some clues in themselves.

P17: Translation part was the easiest one. I could understand the sentences. They were easy for me.

Moreover, the students remarked that vocabulary and listening parts were easy too.

P10: vocabulary part was easy. I knew the words and they were easy.

P21: vocabulary part. I am good at vocabulary. I was able to answer those questions easily.

P2: Listening. I love listening.

P4: Listening was good. I could understand the conversations. They sounded familiar to me.

#### **4.2.7. Participants' final words about the final exam**

To elicit more qualitative data about the final exam the students were asked: Is there anything else that you want to add about the exam? Most participants did not want to add anything. However, with reference to participants' quotations, students hope to have an easier make- up test in case they fail from the final exam. They still thank to their teachers for their effort. Furthermore, in his or her explanation one participant noted that the listening track could have been more clear and understandable.



## **5. DISCUSSION AND CONCLUSION**

The purpose of this study was to examine the multiple choice final exam administered to the non-compulsory preparatory school students at Mehmet Akif Ersoy University (MAKU). The study specifically aimed to analyze the exam in terms of three characteristics of item analysis namely item facility, item discrimination and distractor efficiency.

In this chapter, the results of the present study will be discussed relating the findings with the previous studies presented in the literature review. Moreover, some practical implications for teachers and test developers will be put forward. “The data from item analysis are invaluable tools and should always be followed by a structured discussion” (Ware & Vik, 2009, p. 241). Each research question will be presented in sub-titles and the answers will be given to the questions based on the findings of this study.

### **5.1. What is the difficulty level (item facility) of each item on the final exam test administered to non-compulsory preparatory school students at MAKU?**

The results of the quantitative data 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 to 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. It is important in the sense that items having moderate difficulty levels tend to discriminate better between deep learners and surface learners. Moreover, moderate difficult items are likely to have more functioning distractors.

The results of the present study 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 Olufemi and Oluseyi (2012) those items might not be worth even testing. However, Brown (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, 2004). None the less, 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, I. A., 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 an important content to learn. 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, thereby 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.

From the results it was seen that the complementary relationship between item facility and item discrimination stressed by Gajjar et al. (2014) was also ensured. For instance item # 6 in final exam session 1 was found to be a too easy item for the students. Not surprisingly, that item was not able to discriminate well between high ability and low

ability students and it had a negative ID value. In their study Cechova, et al. (2014) also reported that too easy items did not help to discriminate the test takers having different abilities. Moreover, DiBattista and Kurzawa (2011) also showed that the items having facility values above .90 did not have a satisfactory discrimination index. Similarly, item # 75 and item # 94 in final exam session 1 were found to be too difficult items. These items also could not distinguish surface learners and deep learners and that's why they had negative ID values. In line with the results of the study DiBattista and Kurzawa (2011) also found that the most difficult items had very low discrimination indices and the discrimination index was higher for the items having moderate level of difficulty.

Similar results were found in the second session of the final exam. For instance, item # 6, item # 9, item # 57, and item # 62 in final exam session 2 were found to be too difficult items for the students' levels. These four items also failed to discriminate the high achieving and low achieving students with their negative discrimination indices. The results were consistent with Downing's (2005) who found that flawed items were more difficult and the students could not answer the flawed items although they were able to answer standard items correctly. Similarly, in their study, Olufemi and Oluseyi (2012) also reported that the best discriminating item had a moderate difficulty level like .46. Moreover, Hamzah and Abdullah (2011) found that out of 11 items which were poor in terms of discriminating 10 of them were very difficult items having complex language and 1 of them was a too easy item.

## **5.2. What is the discrimination index (item discrimination) of each item on the final exam test administered to non-compulsory preparatory school students at MAKU?**

The findings of the quantitative data 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 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 negative washback 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; Olufemi & Oluseyi, 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 to improve English language and terminology, peer review, 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).

Oppenheim (2002) stresses the strong relationship between item discrimination and efficient distractors stating that distractors may lure the students having low level of achievement, whereas students having high level of achievements are not deviated by the distractors. From the results it was noted that the relationship between item discrimination and distractor efficiency was supported. For instance, item # 60 in final exam session 1 had negative ID (-0.06) which means the item was not able to discriminate high ability and low ability students. Not surprisingly, this item was answered correctly by 10 (out of 36) high ability students whereas 15 (out of 36) low ability students could answer it correctly. In their study DiBattista and Kurzawa (2011) also found that as the number of functional distractors increased the discriminatory power of both four-option and five-option items improved. In other words, weak



distractors had a dramatic effect on item discrimination and when the number of the weak distractors increased they could even cause the items to lose their discriminatory powers at all.

Similar results were also found in the second session of the final exam. For instance, item # 6 in final exam session 2 had negative ID. That item was answered correctly by 5 (out of 34) high ability students while 10 (out of 34) low ability students answered it correctly. Similarly, item # 17 in final exam session 2 had negative discrimination index. That item was answered correctly more by low ability students rather than high ability students (high ability = 9 (out of 34), low ability = 18 (out of 34)). In their study Tarrant et al. (2009) also showed that there was a strong relationship between the option discrimination and item discrimination. According to the findings of the study items having 3 functional distractors were the most discriminating items. As a result, it was concluded that if an item had a distractor having a high discrimination value, that item was regarded as discriminating overall.

### **5.3. What is the distribution of the response patterns (distractor efficiency) for each of the five options of the items on the final exam test administered to non-compulsory preparatory school students at MAKU like?**

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 group rather than low ability group. These distractors might be even more problematic and they might be discarded or omitted from 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.

Moreover, insufficient time devoted to preparing the exam might result in non-functioning distractors. As a solution to that there might be a specific coordinator ship responsible for preparing the exam, and administering the exam. The coordinator ship might ensure that the choices are independent and they are not overlapping as suggested by Haladyna et al. (2002) (guideline # 22). Moreover, peer review might be avoided because of the reluctance of the instructors to criticize their colleagues. However, Haladyna et al. (2002) suggest that the distractors should base on the typical errors done by the students (guideline # 30). Therefore, the instructors should be willing and free to share their ideas on the students' common mistakes and exchange ideas with each other.

Furthermore, the institution might have no pre-set guidelines to assist the instructors while writing the items (Josefowicz et al. 2002). To construct an effective test with high quality items with effective distractors teachers might benefit from item writing guidelines during modification process (Haladyna, 2004; McDonald, 2007). For instance, instead of a weak distractor "*None of the above (NOTA)*" could be used. Rich and Johnson (1990) note that a NOTA option could be better than a weak distractor and decrease the chance of correct guessing. Their study also proved that items having NOTA option were more discriminating than the usual options not having NOTA option (Rich & Johnson, 1990). Moreover, empirical studies showed that NOTA option did not make the item easier (Crehan et al., 1993; Rich & Johnson, 1990, Frary, 1991). None the less, Haladyna et al. (2002) brings forth that NOTA should be used carefully (guideline # 25).

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 an 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.

#### **5.4. In which part of the exam did the students do well, and in which parts did the students do badly?**

The quantitative results of the study showed that students who took final exam session 1 were most successful in Listening part of the exam (66% success ratio). The results of item facility analysis of the items in Listening part are consistent with this result since most of the items in Listening part have moderate difficulty levels (ranging between .45 and .77). Furthermore, the results also revealed that the students did worst in Dialogue part of the final exam session 1 (32% success ratio). However, item facility analysis of the items in Dialogue part showed that all 4 items had moderate difficulty levels (ranging between .25 and .41).

It is seen that there is a discrepancy between these results. The students' failure in Dialogue part of final exam session 1 might be attributed to their low social skills, or dialogue completion or social conversation skills might not be covered enough during the lessons. More tasks involving reading dialogues, completing dialogues or writing real life situation dialogues might be benefitted. Another reason might be the length of the items. The dialogues might be too long and the students might have lost their concentration on the text. Haladyna et al. (2002) suggest minimizing the amount of reading for each item (guideline # 13).

When it comes to final exam session 2 the students did best in Translation part of the exam (55% success ratio). The results of item facility analysis of the items in Translation part showed that all the items in this part had moderate difficulty levels (ranging between .44 and .56) while one item being too easy ( $p=.85$ ). Therefore, it can be said that the results are consistent. Moreover, the results indicated that the students did worst in Vocabulary part of the final exam session 2 (32% success ratio). The results of item facility analysis of the items in Vocabulary part showed that all the items in this part had moderate difficulty levels (ranging between .16 and .68). Hence, there seems to be a contradiction between the results. The reason of students' fail in

Vocabulary part of the final exam session 2 might be because of not revising or not repeating the new vocabulary. Or, the vocabularies might be over their levels. None the less, Haladyna et al. (2002) bring forth that vocabulary should be kept simple for the test takers (guideline # 8).

### **5.5. Is there a significant relationship between the non-compulsory preparatory schools students' in-year grades point average and the students' final grades point average?**

The results of the quantitative data proved that there is a significant difference ( $p=.000$ ) between non-compulsory preparatory school students' in-year grades point average and their final grades point average ( $p\leq 0.05$ ). The results implied that the students were more successful in their in year midterms or quizzes than the final exam administered at the end of the year. Hence, it can be argued that the final exam was difficult for the students when compared to their in-year grades throughout the year. However, the analysis results showed that most items in final exams had moderate difficulty levels. The students might not have revised the earlier topics enough for the exam. They might be too excited during the exam because it affects 50 percent of their overall score. It could be argued that negative washback could be stimulated for the students.

To reach a more reliable interpretation item analysis of those exams could also be done and more accurate conclusions could be drawn. Such kind of analysis helps teachers construct and revise their items, which is one of the primary aims of item analysis (Cechove et al. 2014; Olufemi & Oluseyi, 2012). Moreover, useful insights and implications for future teaching and assessment process might be drawn from item analysis of the exams (Bodner, 1980; Olufemi & Oluseyi, 2012).

### **5.6. What are the non-compulsory preparatory school students' perceptions about the difficulty level of the items, the discriminatory power of the exam, the efficiency of the distractors and the exam in general administered to the non-compulsory preparatory school students at MAKU?**

The results of the interviews revealed that most of the examinees think that the exam was difficult. Their remarks were contradictory to the quantitative results which showed

that most of the items on the exams had moderate difficulty levels. For discriminatory power of the exam, almost all of the students were of the idea that the exam was discriminating well. Yet, the statistical results showed that most of the items had low discrimination indices. The students might not have a clear perception of the term discrimination or an item's being discriminative.

When it comes to the efficiency of the distractors, most of the students mentioned that the distractors were efficient and strong. Their remarks were parallel to distribution patterns of students' answers. However, there were some non-functioning distractors. Furthermore, the students stated that the exam covered the whole year, both the first and second semesters. That might be the reason of the decrease in students' final grades average when compared to their in year grades. They might have forgotten the vocabulary or grammatical forms covered in the first semester.

## **5.7. Conclusion**

Multiple choice tests have a significant role in higher education. Hence, the findings of the study are particularly of interest in schools or universities where MC exams are widely used as an assessment tool and constitute a significant part of the students' grades. Buckles and Siegfried (2006) state that "the creation of an accessible test bank with high quality, pretested, in-depth multiple choice questions would be of significant service to the profession and to our students" (p.57). To write effective items having high value of item discrimination and of average difficulty with strong distractors the test the teachers should keep item-writing guidelines in mind (Atalmış, 2014). Similarly, Nicol (2007) suggests that if the teachers seek to evaluate the effectiveness of their teaching or the students' performances, they should write quality MC questions and for that they need a pedagogical model or training.

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 are responsible for preparing a practical systematic and reliable test with high validity because important decisions related to students are given according to the exam results. 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 (Olufemi & Oluseyi, 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 to interpret the results of the analysis appropriately.

The instructors who are responsible for preparing the exam should be trained. All test developers should agree on a format and guideline for writing items, and the exam should be prepared advance of the official exam date. A committee should review, critique and approve the content and the format of the exam before it is administered to the students (Josefowicz et al. 2002). Here, it is essential to set the date of the committee well in advance (Wallach, et al. 2006). Moreover, before the committee a peer review among the instructors could also improve the quality of the items and effectiveness of the distractors (Malau-Aduli & Zimitat, 2012).

With formal training and an exam committee, content validity of the items could be increased, technical flaws could be reduced, ambiguity in question interpretations could be detected, item quality could be improved, multiple answers might be picked up, test items might be strengthened in terms of item facility and item discrimination, flawed distractors could be identified and they might be modified (Malau-Aduli & Zimitat, 2012). Such an advance review process might allow test developers to clarify, organize and rewrite their items to increase content validity, reduce technical flaws, and improve overall quality of the exam before administering it to the students (Wallach, et al. 2006).

Therefore, this study was carried out to analyze the multiple choice items in final exams administered to non-compulsory preparatory school students. The data demonstrated that final exams were of high quality in terms of the difficulty levels of the items. The exams had mix of difficulty levels with very limited numbers of too easy or too difficult questions. Since the numbers were limited too easy items could be regarded as warm-up

questions especially for low- ability students, and similarly, too difficult items could be regarded as challenge for high ability students.

The results also showed that the final exams were suffering in terms of discrimination index. There were not found any items having a high difficulty index, and some items had negative discriminaiton indexes. These items should be revised or modified to improve the reliability and validity of the exams. Especially items with negative discrimination index should be analyzed carefully to prevent negative effects of the exams on high-ability students.

As for distractor efficiency although there were found some non-functional disractors they may not create a big problem for the overall quality of the exam. Still, non-functional distractors which have no utility for the quality of the exams should either be modified or omitted from the tests. In that way, students' fatigue and teachers' effort could be diminished, more content could be assessed allowing more items rather than more options, discriminaiton power of the items could be improved.

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.



## REFERENCES

- Ahmadian, M., Yazdani, H., & Darabi, A. (2011). Assessing English Learners' Knowledge of Semantic Prosody through a Corpus-Driven Design of Semantic Prosody Test. *English Language Teaching*, 4(4), 288-298.
- 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.
- 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.
- Babaii, E., & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle?. *System*, 29(2), 209-219.
- Bachman, L. F. (1991). What does language testing have to offer? *Tesol Quarterly*, 25(4), 671-704.
- Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, 25(1), 31-36.
- Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education*, 34(1), 1-15.
- Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher education research & development*, 18(1), 57-75.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of education goals. Handbook I: Cognitive domain (Vol. 1). New York, NY: David McKay Company.
- Bodner, G. M. (1980). Statistical Analysis of Multiple-Choice Exams. *Journal of Chemical Education*, 57(3), 188-90.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, 32(4), 653-675.
- Brown, J. D. (2003). Norm-referenced item analysis (item facility and item discrimination). *Statistics*, 7(2).
- Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. NY: Pearson Education.
- Bruno, J. E., & Dirkwager, 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.

- 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.
- Burton, R. F. (2001). Do Item-discrimination Indices Really Help Us to Improve Our Tests? *Assessment & Evaluation in Higher Education*, 26(3), 213-220.
- 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.
- Burton, R. F. (2005). Multiple - choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1), 65-72.
- Burton, R. F. (2006). Sampling knowledge and understanding: how long should a test be?. *Assessment & Evaluation in Higher Education*, 31(5), 569-582.
- Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher education*, 25(2), 157-163.
- Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education*, 14(4), 398-404.
- Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, 40(2), 218-231.
- Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & cognition*, 43(2), 193-205.
- 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.
- 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.
- 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.
- 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.
- Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, 53(1), 241-247.
- Creswell, J. W., Fetters, M. D., & Ivankova, N. V. (2004). Designing a mixed methods study in primary care. *The Annals of Family Medicine*, 2(1), 7-12.

- 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), 4.
- Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research*, 5(2), 1-17.
- Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10(1), 61-82.
- Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7(3), 235-241.
- 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.
- Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.
- Ebel, R. L. (1967). The relationship of item discrimination to test reliability. *Journal of educational Measurement*, 4(3), 125-128.
- Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, 29(3), 565-570.
- Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, 4(2), 115-124.
- 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.
- Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly*, 11 (1), 69-78.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, 2(1), 37-50.
- 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.
- 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.
- Hamzah, M. S. G., & Abdullah, S. K. (2011). Test Item Analysis: An Educator Professionalism Approach. *Online Submission*.

- Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, 73(2), 94-97.
- Jafarpur, A. (1999). Can the C-test be improved with the classical item analysis? *System*, 27(1), 79-89.
- Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in- house medical school examinations. *Academic Medicine*, 77(2), 156-161.
- Khodadady, E., Pishghadam, R., & Alaei, F. F. (2012). Exploring the Relationship between Social Capitals and English Language Achievement within a Specific Grade and Context. *English Language Teaching*, 5(5), 45-55.
- Kolstad, R. K., Briggs, L. P., & Kolstad, R. A. (1984). The application of item analysis to classroom achievement tests. *Education*, 105(1).
- Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, 211-234.
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & cognition*, 43(1), 14-26.
- Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931.
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical teacher*, 26(8), 709-712.
- Nicol, D. (2007). E- assessment by design: using multiple- choice tests to good effect. *Journal of Further and higher Education*, 31(1), 53-64.
- Nunan, D. (1992). *Research methods in language learning*. Cambridge University Press.
- Oluseyi, A. E., & Olufemi, A. T. (2011). 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.
- 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.
- Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education*, 25(2), 109-119.
- Rich, C. E., & Johanson, G. A. (1990). An Item-Level Analysis of "None of the Above."
- 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.

- Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses?. *Medical Education*, 38(9), 974-979.
- Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35(4), 453-472.
- Shepard, L., & Kirst, M. W. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher*, 20(2), 21-27.
- Simkin, M. G., & Kuechler, W. L. (2005). Multiple- Choice Tests and Student Understanding: What Is the Connection?. *Decision Sciences Journal of Innovative Education*, 3(1), 73-98.
- Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271-286.
- 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.
- Tian, X. (2007). Do assessment methods matter? A sensitivity test. *Assessment & Evaluation in Higher Education*, 32(4), 387-401.
- Toksöz, S., & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, 8(6), 141-146.
- Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1(2), 386-391.
- Vahalia, K. V. Subramaniam, K., Marks, S. C., & De Souza, E. J. (1995). The use of multiple- choice tests in anatomy: Common pitfalls and how to avoid them. *Clinical Anatomy*, 8(1), 61-65.
- Wainer, H. (1988). The future of item analysis. *ETS Research Report Series*, 1988(2).
- 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.
- Walstad, W. B. & Becker, W. E. (1994). Achievement differences on multiple-choice and essay tests in economics. *The American Economic Review*, 84(2), 193-196.
- 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.
- Williams, R. L., & Clark, L. (2004). College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice exams. *Educational Research*, 46(3), 229-239.

- Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, 36(1), 107-122.
- Yonker, J. E. (2011). The relationship of deep and surface study approaches on factual and applied test- bank multiple- choice question performance. *Assessment & Evaluation in Higher Education*, 36(6), 673-686.





**APPENDICES**

## Appendix A. Final Exam (Session I) Used in the Study

### LISTENING

#### *Texting and Driving*

1. The woman is going to \_\_\_\_\_.  
A) the bus station    B) the airport    C) the train station
2. The man receives a text message about \_\_\_\_\_.  
A) a medical appointment    B) a job interview    C) vacation plans
3. The woman wants to \_\_\_\_\_ because he won't stop texting.  
A) hit the man    B) call her brother    C) get out and walk
4. The man and woman are in trouble because \_\_\_\_\_.  
A) they don't have a license    B) the other driver looks scary    C) their window doesn't work
5. The woman suggests \_\_\_\_\_.  
A) catching the bus    B) calling the police    C) talking with the other driver

#### *Class Reunion*

6. What is the woman's name?  
A) Her name is Ashley.    B) It's Amanda    C) The woman's name is Amber.
7. Where is the reunion going to be held?  
A) Mountain Country Club    B) Meadow Country Club    C) Mesa Country Club
8. The reunion starts at \_\_\_\_\_.  
A) 6:00    B) 6:15    C) 6:30
9. How much do two tickets cost for the reunion?  
A) They cost \$20.    B) The price is \$30.    C) The total comes to \$40.
10. Why does James feel uncomfortable about attending the reunion?  
A) James wasn't very good on the football team, and people remember this.  
B) James was rejected by a girl in high school, and he is unsure about seeing her again.  
C) James doesn't know how to dance very well, and he is embarrassed about this.

### GRAMMAR

11. Bronson scored a goal. Yes, a goal \_\_\_\_\_ by Bronson.  
A) is scored    B) is being scored    C) will be scored    D) scored    E) was scored
12. The washing \_\_\_\_\_ by my mother every day for the last twenty years.  
A) is done    B) was doing    C) has been done    D) will be done    E) is being done
13. The next term \_\_\_\_\_ on 16th September.  
A) begin    B) begins    C) began    D) has begun    E) is beginning
14. That ice is dangerously thin now. You \_\_\_\_\_ go ice-skating today.  
A) mustn't    B) might not    C) would mind not to    D) have to    E) don't have to
15. The swimming pool \_\_\_\_\_ at 9 o'clock and \_\_\_\_\_ at 18.30 every day.  
A) is opening / is closing    B) opens / closes  
C) has opened / has closed    D) opened / closed    E) was opening / was closing
16. She caught them while they \_\_\_\_\_.  
A) were talking    B) were talked    C) talked    D) are talking    E) have talked



17. My brother and my sister \_\_\_\_\_ about something when I \_\_\_\_\_ into the room.  
 A) were arguing / were walking    B) argued / walked  
 C) was arguing / was walking    D) were arguing / walked    E) argued/was walking
18. Our teacher, Charlie \_\_\_\_\_ France three times but he doesn't speak French very well.  
 A) visited    B) has visited    C) is visiting    D) will visit    E) has visited
19. Kate has eaten \_\_\_\_\_ lunch already, but I'm saving \_\_\_\_\_ until later.  
 A) her / mine    B) hers / my    C) his / our    D) our / his    E) his / theirs
20. My \_\_\_\_\_ house is over there. He lives just across the street.  
 A) cousin'    B) cousins's    C) cousin's    D) cousins'    E) cousins
21. When I saw \_\_\_\_\_ in the mirror, I could not believe \_\_\_\_\_ eyes. I had turned into a monster. Luckily, it was only a dream.  
 A) myself/its    B) herself /her    C) yourself / your    D) myself / my    E) themselves / their
22. There was \_\_\_\_\_ to help me so I had to do all the cleaning by \_\_\_\_\_.  
 A) somebody / yourself    B) everybody / yourselves  
 C) nobody / myself    D) anybody / yourself    E) nobody / himself
23. If I smoked a cigarette, \_\_\_\_\_ you?  
 A) was it bother you?    B) would it bother you?  
 C) will it bother you?    D) does it bother you?    E) can it bother you?
24. Do you know anyone \_\_\_\_\_ speaks Japanese?  
 A) where    B) which    C) whose    D) who    E) when
25. A: Mary: "I love chocolate."  
 B: Jill: "Mary \_\_\_\_\_ that she \_\_\_\_\_ chocolate."  
 A) says/ loved    B) said / loves    C) say / loving    D) said/ is loving    E) said / loved
26. A: Tom: "I will eat steak for dinner."  
 B: Jack: "Tom \_\_\_\_\_ that he \_\_\_\_\_ eat steak for dinner."  
 A) says/ will    B) said / would    C) say / is going to    D) said/ will    E) said / is going to
- 27.A: What do you know \_\_\_\_ her?  
 B: She's afraid \_\_\_ dogs and she is married \_\_\_ a dentist .  
 A) about / with / with    B) for / of / with    C) about / of / to    D) to / from / by    E) of / with / to
28. I 'm fed up \_\_\_ my sister. She is always angry \_\_\_ me.  
 A) with / for    B) with / with    C) for / to    D) by / with    E) from / to
29. \_\_\_\_\_ I got home , I opened the windows \_\_\_\_\_ it was very hot.  
 A) When / so    B) Before / because    C) When / but    D) While / so    E) When /because
30. \_\_\_\_\_ you finish your homework , you can meet your friends \_\_\_\_\_ don't be late.  
 A) Before / but    B) After / but    C) When / because    D) While / so    E) After / or

### VOCABULARY

31. Capsule hotels are so popular in big cities like Tokyo. They are..... but convenient.  
 A) bright    B) cramped    C) cheap    D) spacious    E) colorful
32. This is a different architecture. It has a steep roof so the rooms are narrow and the \_\_\_\_ are so high.  
 A) balcony    B) gardens    C) window    D) ceilings    E) architects

33. Bats are \_\_\_\_\_ animals. They hunt at night and sleep during the day.  
A) amphibious      B) nocturnal      C) harmless      D) deadly      E) harmful
34. The pie chart on page 190 shows the \_\_\_\_\_ of people who use each type of transport to get to work or school.  
A) diversity      B) results      C) pollution      D) percentage      E) passengers
35. The average healthy man \_\_\_\_\_ about 70 kg.  
A) eats      B) results      C) weighs      D) smells      E) shapes
36. Antalya has a Mediterranean \_\_\_\_\_ with hot and dry summers and mild and rainy winters.  
A) temperature      B) speed      C) cloud      D) rainfall      E) climate
37. Helen loves athletics. She's a big \_\_\_\_\_ of Lionel Messi.  
A) player      B) fan      C) sports      D) ticket      E) game
38. The \_\_\_\_\_ of Germany is Euro.  
A) population      B) region      C) tourism      D) currency      E) religion
39. Scott always knows how to deal with difficult situations. He seems to be a very \_\_\_\_\_ person.  
A) strict      B) level-headed      C) generous      D) short-tempered      E) forgetful
40. We hurried to catch the train but, \_\_\_\_\_, we couldn't catch it.  
A) luckily      B) suddenly      C) strangely      D) unfortunately      E) miraculously
41. After the \_\_\_\_\_, she had an operation to remove some pieces of metal from her legs.  
A) training      B) accident      C) feature      D) refuge      E) mission
42. We are still a small company, but we plan to \_\_\_\_\_ our business over the next year.  
A) advise      B) last      C) launch      D) expand      E) decrease
43. People are angry about the \_\_\_\_\_ in gasoline prices.  
A) fall      B) low      C) rise      D) drop      E) decrease
44. Everyone in the class is expected to \_\_\_\_\_ actively in the competitions  
A) produce      B) predict      C) prove      D) participate      E) protect
45. There aren't enough parking spaces and it's a common \_\_\_\_\_ among the city's residents.  
A) complaint      B) advantage      C) housing      D) signal      E) reason
46. \_\_\_\_\_ means 'very special or different to other things'.  
A) Race      B) Ancient      C) Unique      D) Jockey      E) Tradition
47. This organization gives \_\_\_\_\_ to the students.  
A) grade      B) independent      C) scholarship      D) matter      E) skill
48. The word 'fat' can hurt people. Instead, you can say \_\_\_\_\_.  
A) fun      B) overweight      C) unsuitable      D) healthy      E) educational
49. I don't like her style. Her clothes aren't in style. They look \_\_\_\_\_.  
A) skinny      B) baggy      C) outdated      D) neutral      E) chic
50. She always plans her future. She has very clear \_\_\_\_\_.  
A) directions      B) goals      C) tips      D) areas      E) terms

#### CLOZE TEST

The English word "yoghurt" comes from the Turkish word "yoğurt", (51) \_\_\_\_\_ may be derived from the verb "yogurtmak", meaning "to blend" - a reference to how yoghurt is made. Yoghurt-making involves the introduction of specific kinds of bacteria into pasteurized milk under very carefully controlled temperature and environmental (52) \_\_\_\_\_. Yoghurt is traditionally believed (53) \_\_\_\_\_ by the

Bulgar people of central Asia, although there is evidence of cultured milk products in other cultures as far back as 2000 BC. The earliest yoghurts were probably spontaneously fermented, perhaps by wild bacteria residing inside goatskin bags used for transportation. In Europe, yoghurt remained primarily a food of the central and eastern parts of the continent until the 1900s, when a Russian biologist said (54) \_\_\_\_\_ heavy yoghurt consumption was responsible (55) \_\_\_\_\_ the unusually long lifespans of the Bulgar people. Soon after, yoghurt began to be promoted as a healthy snack, and in 1919 the widespread industrial production of yoghurt in Europe began in Barcelona.

51-A)who B)which C)where D)these E)those

52-A) occupations B) occurrences C) results D) conditions E) disturbances

53- A) invented B) to have invented C) having been invented D) inventing  
E) to be invented

54- A) that B) what C) which D) where E) whose

55- A) to B) among C) over D) for E) through

Houses are buildings that people can live, eat and sleep in. They (56) \_\_\_\_\_ you from dangers and bad weather. Most houses show the lifestyles, traditions and cultures of the people who live in them. Homes and houses have different (57) \_\_\_\_\_ and sizes. They are built of different materials that depend on the climate of the area you live in. Long ago, people (58) \_\_\_\_\_ homes with whatever building materials that they had. In Africa and some islands of the South Pacific they used grass or leaves that grew nearby. In the south-western part of the United States the Pueblo Indians used sun-dried bricks to build (59) \_\_\_\_\_ houses. In the northern part of North America and in northern Europe wood has been the main building material (60) \_\_\_\_\_ many centuries.

56. A) work B) protect C) analyze D) use E) serve

57. A) spices B) shapes C) space D) species E) spare

58. A) are building B) has built C) have built D) built E) build

59. A) their B) they C) themselves D) them E) theirs

60. A) of B) in C) at D) on E) for

#### TRANSLATION

61. You may take your salary in foreign currency if you wish.

- A) Maaşınızı döviz olarak almak istediğinizi belirtiniz.
- B) Maaşınızı döviz olarak alabilmeniz için başvuruda bulunmanız gerekir.
- C) Maaşınızı döviz olarak almak istediğinizi yazı ile bildirin.
- D) Arzu ederseniz, maaşınızı döviz olarak alabilirsiniz.
- E) İsterseniz maaşınızın dilediğiniz miktarı döviz olarak ödenebilir.

62. The parliaments of the member states of the European Union agreed on the use of a single currency in 2002.

- A) Avrupa Birliğine üye devletlerin, kullanımı üzerinde 2002'de anlaştığı tek para birimi vardır.
- B)Avrupa Birliğine üye devletlerin parlamentolarının, kullanımı üzerinde anlaşabildiği tek para birimi 2002'de kabul edilmiştir.
- C)Avrupa Birliğine üye devletlerin parlamentoları, kullanacakları tek para birimini 2002'de kullanma konusunda anlaşmışlardır.
- D)Avrupa Birliğine üye devletlerin parlamentoları, kullanımı üzerinde uzlaştıkları tek para birimini 2002'de piyasa sürmüşlerdir.
- E)Avrupa Birliğine üye devletlerin parlamentoları, 2002'de tek bir para biriminin kullanımı üzerinde anlaştılar.

63. Shakespeare'in döneminde, tiyatro, 1649'da yasaklanıncaya kadar yaygın bir eğlence biçimiydi.

- A) The theatre was a popular form of entertainment from Shakespeare's time and continuing to 1649 when it was banned.
- B) In Shakespeare's time, the theatre was a popular form of entertainment until it was banned in 1649.
- C) The theatre which enjoyed a great deal of popularity in Shakespeare's time was banned in 1649.
- D) In the time of Shakespeare, the theatre was one of the most popular forms of entertainment, but in 1649 it became unpopular.
- E) In 1649 the theatre, which enjoyed so much popularity in the time of Shakespeare, was banned.

64. Günlük yaşamda yaptığımız şeylerin çoğu, alışkanlıklarımızın sonucudur.

- A) We do certain things in our everyday life so often that they become our habits.
- B) Most of the things that we do in everyday life are the result of our habits.
- C) Most of our routines come from the fact that we do some things every day.
- D) Habitually we do lots of things in our everyday lives.
- E) Most part of our everyday living is made up of the things we do often.

65. We'll need a few more workers to meet the increasing demand.

- A) İşçileri karşılaması için birini göndermemiz gerekiyor.
- B) İşçi sayımızı arttırmadıkça, talebi karşılayamayız.
- C) Artan talebi karşılamak için birkaç işçiye daha ihtiyacımız olacak.
- D) İşçilerin bazı taleplerini karşılamamız gerekiyor.
- E) Talepteki artış daha fazla işçi çalıştırmamızı gerektiriyor.

66. İtiraf etmeliyim ki iş gereği yurt dışına çıkmak benim için ilginç bir deneyim olacak.

- A) I have to admit that it will be necessary for me to go abroad to become experienced in my job.
- B) I must confess that it will be an interesting experience for me to go abroad on business.
- C) I'm interested in going abroad on business to become experienced in my job.
- D) I claim that going abroad on business will be an interesting experience for me.
- E) I must admit that going abroad on business was an interesting experience for me.

### READING

The printing press was invented by Gutenberg in the city of Mainz, in Germany. He built and operated the printing press with movable metal letters. In fact, simple printing methods had existed for centuries; however "they" had to be done by hand and took a long time. What made Gutenberg's press so different was that the individual letters themselves could easily be moved to create different pages. That made it possible to print entire books more cheaply and more quickly than ever before.

67. What did Gutenberg invent?

- A) Gutenberg alphabet.      B) Handwritten books.      C) The printing press.      D) Metal letters.
- E) Simple printing methods.

68. The basic new feature of Gutenberg's printing press \_\_\_\_\_.

- A) was that all the pages of a book were printed at the same time
- B) was that it could easily be operated by unskilled workmen
- C) was that the printing of books was less costly although it took a long time to do
- D) was the use of metal letters that could be moved into different positions
- E) made it possible to print books without any error at all

69. Actually, the history of printing \_\_\_\_\_.

- A) first begins with Gutenberg's invention
- B) has always been associated with Germany
- C) begins before the time of Gutenberg
- D) runs parallel to the history of books
- E) gives less importance to Gutenberg's invention than it deserves

70. Gutenberg's printing techniques \_\_\_\_\_.
- A) made printing more complicated and time-consuming
  - B) was not as important as it has often been thought
  - C) was not used outside Germany for a long time
  - D) speeded up the printing of books
  - E) adopted the metal letters system of easier printing methods

71. What does "they" refer to in the paragraph?
- A) Metal letters.
  - B) Centuries.
  - C) Individual letters
  - D) Books that are printed.
  - E) Simple printing methods.

Hector Munro was born in Burma. He was the son of a police officer, Charles Munro. He was married to Mary Mercer, the daughter of Admiral Samuel Mercer. Her nephew, Cecil Mercer, later became a famous novelist. Hector's mother died when "he" was two, and then he was sent home to Scotland to live with relatives. His formal education ended with grammar school, but his father taught him on travels. In 1893, his father found him a job in the Burma police, but his health forced his return to Britain. There he took up a career in writing, and it was while doing political sketches for The Westminster Gazette that he used the penname of Saki. After serving for a time as a foreign reporter for The Morning Post, he returned to London to write stories and novels. When World War I began, he became an ordinary soldier in the army and unfortunately died in action in 1916.

72. Munro's father \_\_\_\_\_.
- A) helped his son's education
  - B) did not want his son to work in Burma
  - C) was one of the founders of the Burmese police force
  - D) was a great traveler himself and encouraged his son to follow his example
  - E) tried to persuade his son not to join the army in World War I

73. Munro \_\_\_\_\_.
- A) commanded in World War I
  - B) adapted himself well to the climatic conditions of Burma
  - C) disappointed his father with his decision to return to England
  - D) was both a journalist and a writer
  - E) chose the name "Saki" because it was an easy name for his readers to remember

74. Before he became a writer, Munro \_\_\_\_\_.
- A) worked as a journalist to cover events of World War I
  - B) spent all his time in Scotland with relatives
  - C) travelled very little, but read a lot
  - D) enjoyed good health
  - E) served, for some time, as a policeman

75. Samuel Mercer was Hector's \_\_\_\_\_.
- A) father
  - B) mother
  - C) grandfather
  - D) nephew
  - E) cousin

76. The word "he" refers to \_\_\_\_\_.
- A) Cecil Mercer
  - B) Hector Munro
  - C) Samuel Mercer
  - D) Charles Munro
  - E) Mary Mercer

The Toowoomba Carnival of Flowers is the most important floral event in Queensland, Australia. It takes place every year in late September and it lasts 10 days. Lots of people attend it, so last year I decided to go and see what it was all about. The carnival began with a competition for the best garden in town. After that, there was a spectacular street parade of convertible cars covered in flowers. Then followed a procession of pipers, dressed in "traditional" Scottish tartan kilts and groups of dancers, wearing bright costumes. Finally, there was the Flower Queen contest, in which people voted for the most beautiful girl of the day. I enjoyed every second of this cheerful floral feast. I felt relaxed walking through the beautiful parks and gardens. I was stunned by the street parade and I had a lot of fun voting for the queen in the

contest. On the whole, it was an amazing event. As I was leaving Toowoomba, I felt really sad that the event was over. I couldn't help thinking how much fun I had. So, if you ever plan a trip to Toowoomba, try not to miss the lively events and the cheerful atmosphere of this exciting festival.

77. With which floral event does the Carnival start?

- A) A floral queen contest.
- B) Spectacular street parades with convertible cars covered with flowers.
- C) Competition of the best garden in town.
- D) A procession of pipers dressed in traditional Scottish tartan kilts.
- E) Dancers wearing colorful bright costumes.

78. What does "traditional" mean?

- A) Classical
- B) Fresh
- C) Unusual
- D) New
- E) Modern

79. Which of the following is not mentioned in the passage?

- A) Flower queen contest.
- B) Flower carnival in Queensland.
- C) The traditional food of Queensland.
- D) Procession of pipers dressed in traditional clothes.
- E) Spectacular street parades.

80. Why did the author feel sad while leaving Toowoomba?

- A) His candidate wasn't chosen in the queen contest.
- B) He didn't like the traditional Scottish kilts.
- C) The event was over and he had to leave.
- D) Walking through the parks were exciting
- E) The best garden which was chosen wasn't his garden.

81. What does the writer advice the readers about the trip to Toowoomba?

- A) Visit there only once.
- B) Never go to Toowoomba.
- C) Eat traditional food of Toowoomba.
- D) Don't miss to attend to floral festival.
- E) Wear traditional Scottish tartan kits.

If you visit Japan, you might choose to travel by Shinkansen train. These high speed trains connect the major cities of Japan. They are nicknamed "bullet trains" because they go very fast and have pointy noses like a bullet. They are very punctual, often leaving on time. They are also comfortable. Most importantly, bullet trains are very safe. In their 35-year history, there have been only a few accidents and no deaths. The only disadvantage of "them" is that they are expensive. A ticket to travel to another city will be "almost" the same with an airline ticket. However, if you fly, you will land at an airport far away from the city. Train stations are usually right in the middle of a city.

82. These trains are called "bullet trains" because of their \_\_\_\_\_.

- A) safety and shape
- B) safety and timing
- C) speed and shape
- D) speed and timing
- E) safety and speed

83. According to the passage, what is the most important fact about bullet trains?

- A) They have plenty of leg room
- B) They are punctual
- C) They are comfortable
- D) They are very safe
- E) They aren't expensive

84. What is the disadvantage of the bullet trains

- A) They are dangerous to travel in
- B) They are expensive
- C) They have no leg rooms
- D) They travel slowly
- E) Their stations are so far away from the city

85. What does "almost" mean in the paragraph?

- A) exactly
- B) maybe
- C) nearly
- D) definitely
- E) probably

86. What does "them" refer to in the paragraph?

- A) accidents
- B) bullet trains
- C) tickets
- D) stations
- E) seats

## DIALOG COMPLETION & SITUATION

87. Martin : Have you seen my glasses anywhere?  
I seem to have lost them.

Fiona : \_\_\_\_\_

Martin : Am I! How silly of me!

- A) Try looking in the mirror. You're wearing them!
- B) Yes, they're in the bathroom, on the washbasin.
- C) Oh, you haven't lost them again, have you?
- D) You really should be more careful with your belongings.
- E) No, I haven't, but I'll help you look for them.

88. Maggie: What a beautiful dress. Where did you get it?

Wendy: Well, actually, I had it specially made for me.

Maggie: \_\_\_\_\_

Wendy: No, not really. The designer is a friend of my mother's.

- A) Wow! That must have cost a lot of money.
- B) Oh really! How can you afford to buy exclusive clothes?
- C) Never! It looks just like the one Rita was wearing on Saturday.
- D) I must say it's been made very well —just your style.
- E) I wish I could have clothes made for me sometime

89. Joe : Officer, something in my car has been  
stolen!

Policeman : \_\_\_\_\_

Joe : My briefcase, and the stereo.

Policeman : You'd better come down to the  
station and file a report.

- A) Do you know what was taken from your car, sir?
- B) Is there anything else you'd like to report, sir?
- C) Did the thieves hurt you sir?
- D) Can you give us an accurate description of your car, sir?
- E) Did you lock the door properly when you left it sir?

90. Donna: Are you planning to go to the graduation party?

Jill : \_\_\_\_\_

Donna: You shouldn't miss it. This is something that cannot be repeated in your life.

Jill : You may be right, in fact. I'd better think about it.

- A) Sure, I have been waiting for this day for four years.
- B) I haven't decided yet. What do you think?
- C) They say there will be around 2000 people there.
- D) That's a good idea. Where will it take place?
- E) Absolutely. Most of my close friends will be there.

91. A friend of yours is taking an exam tomorrow and he's worried about it. Try to make him feel relaxed:

- A) Have I ever told you how terrible my last exam was?
- B) If I were you, I would feel the same thing.
- C) If you ever miss that chance, I don't think there will be some others.
- D) You needn't stress about such an easy exam. I'm sure you will pass it.
- E) I think you didn't study enough. Anyway, I will get a better grade than you.

92. Your old neighbor is in a nostalgic mood, but you have to leave for work. You want to escape from the next story without being rude, so you say:

- A) I've had enough of your old stories.
- B) I will visit you as soon as I return from work to discuss this.
- C) This is all very interesting but it was before my time.
- D) Oh yes, I remember that. I used to work near there.
- E) I'm terribly sorry but I must go or I shall be late for work.

93. You are at the cinema with a friend and it's full. A woman is sitting in front of you wearing a huge hat which blocks your view of the screen. You don't want to cause an argument but you really can't see anything, so you say:

- A) Take that hat off immediately! You should be more considerate.
- B) Let's go and sit in one of those free seats over there.
- C) You ought to sit somewhere else if you want to wear that hat.
- D) This is an excellent film. I'm so glad we came to see it.
- E) I'm terribly sorry but could you possibly take your hat off, please?

94. You're in a cafe with a friend. Every time you go there, the waiter gets your order wrong. This time you have ordered an orange juice and a cup of tea. He brings over a cake and two coffees. You complain about this and say:

- A) Thank you very much for your help. It looks lovely.
- B) Excuse me, I wonder if you'd mind bringing us something to eat?
- C) This is ridiculous. You've brought the wrong thing again!
- D) Thank you, but we didn't order a cake.
- E) An orange juice and a cup of tea please, as usual.

95. Your sister wants to go to the cinema tonight. However, you are too tired to go out. You say politely trying not to hurt her:

- A) I really think we should stay at home.
- B) No, there isn't a good film on at the moment.
- C) Would you mind if we postponed it to another night because I'm feeling exhausted.
- D) Are you joking? You didn't buy the tickets, did you?
- E) I think we should hurry up, the film starts in an hour.



## Appendix B. Final Exam (Session II) Used in the Study

### LISTENING

#### *Airline Safety*

1. What is the first item that the man has in his carry-on bag?  
A) water      B) medication      C) a large bottle of shampoo
2. The man is carrying a lighter because he \_\_\_\_\_.  
A) enjoys smoking cigarettes      B) is worried about his safety      C) he forgot to take it out of his bag
3. The next illegal item the man has with him is \_\_\_\_\_.  
A) a live snake      B) Firecrackers      C) a huge knife
4. The man received this item from \_\_\_\_\_.  
A) a close friend      B) a relative      C) a stranger
5. The man doesn't know the airline's rules because \_\_\_\_\_.  
A) the sign was written in Chinese      B) he didn't see the sign      C) he wasn't paying careful attention

#### *Hotel Reservations*

6. The man makes a reservation finally for which day?  
A) March 20<sup>th</sup>      B) March 21<sup>st</sup>      C) March 22<sup>nd</sup>
7. Why doesn't he want to reserve the suite?  
A) It doesn't have a nice view.      B) It doesn't come with a sauna bath.      C) It's too expensive.
8. What kind of room does the man prefer?  
A) a non-smoking room      B) a smoking room      C) either one is okay
9. Including tax, how much is the man's room?  
A) 80 dollars      B) 88 dollars      C) 96 dollars
10. How do you spell the man's name?  
A) Maxner      B) Maexner      C) Mexner Grammar

### GRAMMAR

11. The essays \_\_\_\_\_ on Monday.  
A) have to be get in      B) have to be give in      C) have to be have in      D) have to be took in  
E) have to be handed in
12. The task \_\_\_\_\_ into smaller, manageable mini-tasks.  
A) is brake down      B) is braking down      C) is broke down      D) is broken down      E) is broking down
13. I will call you when I \_\_\_\_\_ back.  
A) will come      B) comes      C) come      D) have come      E) came
14. It's the law. They \_\_\_\_\_ have a blood test before they get married.  
A) might      B) could      C) should      D) may      E) have to
15. A: Mary: "I can't find a job."  
    B: Jill: "Mary \_\_\_\_\_ that she \_\_\_\_\_ find a job"  
A) says/ could      B) said / couldn't      C) say / can't      D) said/ could      E) said / can
16. A: Mary: "I don't like spinach."  
    B: Jill: "Mary \_\_\_\_\_ me that she \_\_\_\_\_ like spinach."  
A) tells/ likes      B) told/ didn't like      C) tell/ didn't like      D) told / don't like  
E) tell / doesn't like

17. Mike was sitting \_\_\_\_ his desk when Bill called; Bill was \_\_\_\_ Hawaii \_\_\_\_ holiday.  
 A) at / in / in      B) on / in / on      C) at / in / on      D) in / on / at      E) in / at / on
18. A: I've lost my keys again.  
 B: I will help you look \_\_\_\_ them.  
 A: That's very kind \_\_\_\_ you.  
 A) for / about      B) at / for      C) with / for      D) for / of      E) about / by
19. \_\_\_\_ you cross the road, always look both sides \_\_\_\_ be careful  
 A) Before / and      B) After / and      C) Before / but      D) While / because      E) When / but
20. \_\_\_\_ I went out, it was raining \_\_\_\_ I opened my umbrella.  
 A) When / because      B) When / so      C) Before / or      D) While / but      E) After / because
21. We usually \_\_\_\_ vegetables in our garden but this year we \_\_\_\_ any.  
 A) are growing / don't grow      B) grew / haven't grown      C) grow / aren't growing  
 D) grow / don't grow      E) are growing / aren't growing
22. Harry \_\_\_\_ and we \_\_\_\_ him the news. \_\_\_\_  
 A) woke up / were telling      B) woke up / told      C) was waking up / have told  
 D) wakes up / told      E) has woken up / were telling
23. At six o'clock this morning most of us were asleep in bed. But Atheer \_\_\_\_ for today's grammar.  
 A) studied      B) studies      C) was studying      D) has studied      E) is studying
24. I don't know what the road is like now because I \_\_\_\_ the place for twenty years.  
 A) didn't see      B) wasn't seeing      C) won't see      D) don't see      E) haven't seen
25. A: UFOs don't exist, so you cannot have seen one.  
 B: I tell you I saw \_\_\_\_ with \_\_\_\_ own eyes.  
 A) them / my      B) theirs / our      C) they / mine      D) themselves / ours      E) their / me
26. The calculator is old. Some of \_\_\_\_ are broken.  
 A) its keys      B) it keys'      C) it's keys      D) its ' keys      E) it key's
27. The old woman lived alone, with \_\_\_\_ to look after \_\_\_\_.  
 A) anyone / hers      B) anyone / hers      C) somebody / her      D) no one / her      E) everyone / her
28. \_\_\_\_ in the village went to the party but \_\_\_\_ enjoyed it very much.  
 A) Someone / no one      B) Everyone / anyone      C) No one / everyone      D) Anyone / someone  
 E) Everyone / no one
29. If your boss asked you to take an extra work without more money, \_\_\_\_.  
 A) will you agree to do so?      B) can you agree to do so?      C) would you agree to do so?  
 D) did you agree to do so?      E) do you agree to do so?
30. The train \_\_\_\_ goes to Madrid leaves from platform 2.  
 A) when      B) who      C) whose      D) where      E) which
- VOCABULARY**
31. My sister \_\_\_\_ me to become a manager, just like her.  
 A) achieved      B) proved      C) inspired      D) reached      E) existed
32. Many works of art were \_\_\_\_ in the fire.  
 A) caused      B) increased      C) respected      D) damaged      E) predicted
33. She \_\_\_\_ the other runners and went on to win the race.  
 A) protected      B) improved      C) taught      D) overtook      E) produced

34. On her first day at work, her \_\_\_\_\_ did their best to make her feel welcome.  
A) colleagues      B) customers      C) graduates      D) experts      E) researchers
35. Studying abroad provides a great \_\_\_\_\_ to learn a foreign language.  
A) professional      B) opportunity      C) talent      D) pattern      E) fact
36. When you are going on a volunteer travel to a foreign country, it's important for health to get necessary\_\_\_\_\_  
A) chocolate      B) sandals      C) vaccinations      D) money      E) backpack
37. The Siberian tiger is a/an \_\_\_\_\_ mammal. Few of them are found in the wild and three tiger types are already extinct.  
A) endangered      B) average      C) habitat      D) reptile      E) venomous
38. 10 million people \_\_\_\_\_ from Sapporo to Tokyo every year. They usually take the bullet train which travels at up to 300 kph.  
A) drive      B) commute      C) fly      D) prefer      E) work
39. The kiwi is the most \_\_\_\_\_ bird in the world. Male and female kiwis live together for 30 years.  
A) unique      B) romantic      C) nocturnal      D) usual      E) powerful
40. Platypus is a strange Australian animal. It is a mammal but it \_\_\_\_\_eggs.  
A) lays      B) catches      C) lives on      D) looks for      E) hunts
41. South Africa is a dry country. It has an average \_\_\_\_\_ of about 464 mm a year.  
A) temperature      B) speed      C) cloud      D) rainfall      E) climate
42. Cricket is a slow game. I think it's \_\_\_\_\_ .  
A) boring      B) exciting      C) safe      D) dangerous      E) different
43. In the summer, we go mountain climbing. We climb up to the \_\_\_\_\_ .  
A) beach      B) river      C) field      D) valley      E) hill
44. Jennifer did an \_\_\_\_\_ job on her report. That's why she got a raise.  
A) ridiculous      B) disgusting      C) absurd      D) outstanding      E) weird
45. I can't believe I locked myself out of my house. I feel so \_\_\_\_\_ .  
A) marvelous      B) disgusting      C) dumb      D) fabulous      E) unusual
46. If you \_\_\_\_\_ something, you put it somewhere high.  
A) stick      B) celebrate      C) take part      D) succeed      E) hang
47. Carlos is very good at thinking of new ideas and making interesting things. He is really \_\_\_\_\_ .  
A) creative      B) professional      C) similar      D) boring      E) natural
48. The verb \_\_\_\_\_ means 'design or make something new'.  
A) protect      B) invent      C) develop      D) type      E) keep
49. Jack doesn't make much money. His \_\_\_\_\_ is low.  
A) server      B) work      C) pleased      D) cost      E) wage
50. We don't have enough money to buy that house. We should get a/an \_\_\_\_\_ from the bank.  
A) loan      B) savings      C) salary      D) account      E) tip

### CLOZE TEST

The Romantic Age in England was part of a movement that affected all the countries of the Western World. There were (51) \_\_\_\_\_ many forms of romanticism so it is difficult (52) \_\_\_\_\_ of the movement as a whole. It tended to align itself with the humanitarian spirit of the democratic revolutionaries. (53) \_\_\_\_\_ romantics were not always democrats and democrats were not always revolutionaries. Perhaps

the (54) \_\_\_\_\_ thing to say is that romanticism was an attempt to discover the mystery (55) \_\_\_\_\_ the World.

51. A) too      B) few      C) either      D) neither      E) enough  
52. A) speak      B) for speaking      C) spoke      D) to speak      E) speaking  
53. A) However      B) Since      C) Although      D) Because      E) While  
54. A) safe      B) safest      C) safety      D) safely      E) safer  
55. A) at      B) for      C) of      D) from      E) to

Can we see (56) \_\_\_\_\_ the earth is a globe? Yes, we can when we watch a ship that sails out to sea. If we watch closely, we see that the ship begins (57) \_\_\_\_\_. First the bottom of the ship disappears, and then the ship seems to sink lower and lower (58) \_\_\_\_\_, we can only see the top of the ship, and then we see nothing at all. What is hiding the ship from us? It is the earth. Stick a pin most of the way into an orange, and (59) \_\_\_\_\_ turn the orange away from you. You will see the pin disappear, (60) \_\_\_\_\_ a ship does on the earth.

56. A) or      B) where      C) that      D) when      E) whether  
57. A) being disappeared      B) to be disappeared      C) to have disappeared      D) to disappear  
E) having disappeared  
58. A) until      B) after that      C) since      D) by the time      E) unless  
59. A) reluctantly      B) accidentally      C) slowly      D) fast      E) suddenly  
60. A) and      B) however      C) just      like      D) by the way      E) similar

### READING

For most people, being a member of a large family is sometimes hard. Usually there isn't enough money, so everyone has to do various things. There are, however, certain advantages; in fact, there are probably more advantages than disadvantages. One day, I saw a family setting off on a day out. The parents who looked young "themselves" were carrying lots of bags. The biggest child who was perhaps fifteen carried a football. His sister who was perhaps two years younger carried the family lunch. The four smaller children also had things to carry. The youngest of them carried a toybear that was almost as big as herself. The family was catching a bus and looked very happy. I wished I could have gone with them where they were going.

61. The family members \_\_\_\_\_  
A) didn't use to go out for the day like this  
B) very rarely have a day out together  
C) seldom take a bus at weekends  
D) are clearly a very rich one  
E) know how to share their jobs
62. Although these parents have many children, \_\_\_\_\_  
A) they spend very little time with them  
B) they don't really seem to care about them  
C) they don't want to spend much money on them  
D) it seems that life has not aged them  
E) they like the little one most

63. From the passage we can conclude that the writer \_\_\_\_\_.
- A) himself comes from a large family
  - B) is very critical of large families
  - C) is more interested in the parents than in the children
  - D) feels sorry because all the children have got things to carry
  - E) seems to like large families

64. In total, there are \_\_\_\_\_ people in this family.
- A) five
  - B) six
  - C) seven
  - D) eight
  - E) nine

65. The word "themselves" refers to \_\_\_\_\_.
- A) the parents
  - B) the sons
  - C) the young children
  - D) the children
  - E) the daughters

New Guinea is home to some of the world's strangest creatures. For instance, there is a special type of kangaroo that lives in trees. There are also lizards that are five meters long, and butterflies that are as big as dinner plates. New Guinea is an island and it is almost as big as the state of Texas, but it has as many bird species as North America does. One reason can be that it has remained "isolated" from the rest of the world. "It" has had no contact with the other parts of the world. One another reason is that it has an incredible variety of ecological characteristics. It has tropical rain forests, glaciers and other kinds of ecological characteristics.

66. New Guinea \_\_\_\_\_.
- A) has few bird species
  - B) is very similar to Texas
  - C) has different ecological characteristics
  - D) is a place where you cannot see strange animals
  - E) is increasing its contact with North America

67. Kangaroos that live in trees \_\_\_\_\_.
- A) can be seen in every part of the world
  - B) live only in rain forests
  - C) are smaller than an average kangaroo
  - D) are just one example of strange animals in New Guinea
  - E) like eating butterflies

68. A good title for this passage would be \_\_\_\_\_.
- A) The People in New Guinea
  - B) The Strange Animals in New Guinea
  - C) The Glaciers in New Guinea
  - D) New Guinea and Texas
  - E) New Guinea's Location

69. "Isolated" means \_\_\_\_\_.
- A) very cold
  - B) similar to
  - C) different from
  - D) very close to
  - E) far away

70. What does "it" refer to?
- A) Texas
  - B) North America
  - C) New Guinea
  - D) World
  - E) Rain Forest

Fifty years ago, when I was a child, photographs were not of general interest. Photographs were taken on special occasions, at weddings and on birthdays, for instance. These pictures were usually kept in a box and brought out time to time to show the family. Nowadays photography is regarded as an art, just as painting is. Many photographic exhibitions are held and there are many magazines which "deal with" the art of photography.

71. When the writer was young, \_\_\_\_\_.
- A) he was very interested in photography
  - B) people didn't think of photography as an art
  - C) he always took photographs on his birthdays
  - D) people used to go to photographic exhibitions
  - E) he took a lot of family photographs

72. How did the people keep their photographs in the past?

- A) They kept them in photo frames.                      B) They kept them in a box.  
C) They put them under the carpet.                      D) People used to show them to their families.  
E) They wouldn't take photographs.

73. During recent years, \_\_\_\_\_.

- A) photography has become a popular form of art  
B) a lot of people took photographs of good paintings  
C) photography has stopped being an art  
D) photographic exhibitions are often advertised in magazines  
E) more and more people take photographs at weddings

74. The passage compares \_\_\_\_\_.

- A) public interest in painting today and fifty years ago  
B) photographic exhibitions and painting exhibitions  
C) wedding photographs and birthday photographs  
D) photography today and photography fifty years ago  
E) family photographs today and fifty years ago

75. What does "deal with" mean in the paragraph?

- A) put away                      B) are interested in                      C) go on                      D) take off                      E) get away

All firms spend a great deal of money on advertising their goods, and when we buy these goods we have to pay extra to cover the cost of advertisements. Still, most of us get a certain amount of "pleasure" out of advertisements "themselves", especially out of the ones on the radio and the television. Further, newspapers and magazines are sold to us cheaply because publishers collect a lot of money from advertisers.

76. Advertisements are \_\_\_\_\_.

- A) often entertaining.                      B) expensive and useless.                      C) sold to us cheaply.  
D) published only in newspapers.                      E) useless.

77. All firms \_\_\_\_\_.

- A) advertise on the radio.                      B) sell their advertisements cheaply to magazines.  
C) should be banned from advertising on the television.  
D) do not wish to spend any money on advertising  
E) pay a certain amount of money for advertisement

78. Newspapers \_\_\_\_\_.

- A) are cheaper than magazines.                      B) are published by advertisers.  
C) would cost more if they did not print advertisements.                      D) are advertised on the radio.  
E) are sold more than magazines

79. What does "pleasure" mean in the paragraph ?

- A) gladness                      B) worry                      C) sadness                      D) hope                      E) fear

80. What does “themselves” refer to in the paragraph?

- A) publishers      B) goods      C) newspapers      D) advertisements      E) magazines

#### TRANSLATION

81. Faiz oranlarındaki artışa rağmen kardeşim bankadaki hesabını kapattı; ben de öyle yaptım.

- A) Despite the rise in interest rates, my brother has taken money from his bank account, and so have I.  
B) Despite my brother closed his account at the bank, I did too, even though interest rates are high.  
C) Despite interest rates have risen, both my brother and I have opened an account at the bank.  
D) Despite a rise in interest rates my brother would have closed his account at the bank and so would I.  
E) Despite the rise in the interest rates, my brother has closed his account in the bank, and so have I.

82. Can you tell me exactly how I can get to the post office?

- A) Mektubu postaneden nasıl alabileceğimi bana açıkça söyley misin?  
B) Postaneye nasıl gidebileceğimi bana tam olarak söyleyebilir misiniz?  
C) Mektupla postaneye nasıl gidileceğini bana iyice söyley misiniz?  
D) Postanede nasıl iş bulabileceğimi bana ayrıntılarıyla açıklar mısınız?  
E) Mektubu almak için postaneye nasıl temas kuracağımı açıkça söyley misiniz?

83. Women drivers usually prefer small cars because they are easy to park.

- A) Park etmekte zorlanan sürücülerin, çoğunlukla küçük arabalar tercih etmeleri gerekir.  
B) Park etmesi kolay olduğu için, küçük arabaların çoğu bayan sürücüler tarafından satın alınmaktadır.  
C) Bayan sürücüler park etmesi kolay olduğu için genellikle küçük arabaları tercih ediyorlar.  
D) Küçük arabaları park etmek daha kolay olduğu için, bayan sürücülerin tercihi bu yönde olmalıdır.  
E) Bayan sürücüler tarafından kullanılan küçük arabaları park etmek oldukça kolaydır.

84. Eleştirmenler, kitap okumanın, okuyucu ile yazar arasında bir tür sohbet olması gerektiğini her zaman söylemişlerdir.

- A) Critics always tell us to read a book like we were having a conversation with the writer.  
B) Critics are always saying the idea that reading a book is like having a conversation with the writer.  
C) According to some critics, we should always see reading as a conversation between the writer of the book and the reader.  
D) Critics always tell us that reading is a conversation between a writer and a reader.  
E) Critics have always said that reading a book should be a kind of conversation between the reader and the writer.

85. The Suez Canal, which connects the Mediterranean and the Red Sea, was designed and built by the French engineer De Lesseps.

- A) Akdeniz ile Kızıl Deniz’i birleştiren Süveyş Kanalı, Fransız mühendis De Lesseps tarafından tasarlanmış ve inşa edilmiştir.  
B) Fransız mühendis De Lesseps’in tasarlamış ve inşa etmiş olduğu Süveyş Kanalı, Akdeniz ile Kızıl Deniz’i birleştirir.  
C) Süveyş Kanalı’nı tasarlayıp inşa eden Fransız mühendis De Lesseps, Akdeniz ile Kızıl Deniz’i birleştirmeyi amaçlamıştır.  
D) Süveyş Kanalı’nın Fransız mühendis De Lesseps tarafından tasarlanıp inşa edilmesiyle, Akdeniz’in Kızıl Deniz ile birleşmesi sağlanmıştır.  
E) Akdeniz ile Kızıl Deniz’in birleşmesi, Fransız mühendis De Lesseps’in Süveyş Kanalı’nı tasarlayıp inşa etmesiyle olmuştur.

86. Uzmanlar, dünyanın yiyecek üretiminin nüfus büyümesiyle aynı oranda artmadığını düşünüyorlar.

- A) It's thought by the experts that the increase in food production isn't at the same rate as in population growth.  
B) Experts think that food production in the world isn't increasing at the same rate as population growth.  
C) According to the experts, food production can't keep pace with the population growth.  
D) The population of the world is growing at such a high rate that experts are trying to find out how to increase food production equally.  
E) Experts think that food isn't yet produced at such a rate as to meet the requirements of the growing population.

### DIALOG COMPLETION & SITUATION

87. Estate Agent: What sort of house are you looking for, Mr. Reynolds?

Mr. Reynolds : We'd like a two-storey house, with three or four bedrooms and a separate garage.

Estate Agent : \_\_\_\_\_

Mr. Reynolds : Oh yes, because we've got two children and a dog.

- A) What sort of car do you have?
- B) Would you like it to have a garden as well?
- C) Houses like that are quite difficult to find these days.
- D) And how much were you thinking of paying?
- E) How far from the city-centre would you like to be?

88. Policeman : Can you give me a description of the man who was running out of the bank?

Witness: Well, he was about 25 years old; longish fair hair and he was wearing jeans and a blue not blue — green jacket.

Policeman: \_\_\_\_\_

Witness : I am.

- A) Was he carrying a gun or a knife?
- B) So you think he was about 25 years old, then?
- C) Did he have any other distinguishing features?
- D) Are you absolutely certain it was green?
- E) Would you recognize him if you saw him again?

89. Your teacher gives you back some homework. You look at it and are fairly certain that the teacher has made a mistake in the marking and that you were right. So you say:

- A) I really didn't expect to get such a good grade for this.
- B) That homework was a lot easier than I expected.
- C) Could you have another look at my paper?
- D) Excuse me, but I think you've missed one of my mistakes.
- E) Why do you always make such ridiculous mistakes?

90. You are at a party when a woman accidentally spills red wine down your new white dress. You know the dirt probably won't come out. The woman is so apologetic and embarrassed that you feel a little sorry for her. You say to her:

- A) You're so dumb. Look what you've done!
- B) Don't worry! It really doesn't matter at all.
- C) This dress was new and now you've ruined it.
- D) You know you'll have to pay for the cleaning!
- E) Oh, I'm so sorry! How careless of me.

91. Shop assistant: How can I help you?

Customer: I'm looking for a book on Greek mythology.

Shop assistant: Do you know the author of the book?

Customer: \_\_\_\_\_.

Shop assistant: Is it D'aularie?

Customer: Yes! Ah! D'aularie. Thank you.

- A) I think it's D'aularie's book of Greek Myths.
- B) Mythology is something that people aren't interested in very much, isn't it?
- C) I really can't decide. Can you show me all the books on mythology?
- D) Oh, it was something like Dulerie.
- E) If you don't have that book, I can have another writer's.



92. Paul : Let's eat out tonight. Where would you like to go?

Sue : \_\_\_\_\_.

Paul : Good idea! Which one?

Sue : The Chinese one.

- A) You choose. You eat out more than I do.
- B) Anywhere, so long as the food is good.
- C) Will there be time for a quick snack before the film starts?
- D) Well, certainly not that place you took me to last week.
- E) Shall we try one of those new foreign restaurants?

93. You are in Vienna, Austria. Someone asks you where the Opera Building is. You don't know exactly where it is. You say:

- A) It is near the Theater Building, next to the National Gallery.
- B) Sorry, but I can't help you. I am a foreigner here.
- C) I don't think there is such a place as the Opera Building around here.
- D) Go along the street and turn left. You will see the big building.
- E) Why don't you ask the Post Office?

94. You are at a restaurant abroad. You are a vegetarian, but the menu of the restaurant mostly consists of meat meals. You ask the waiter:

- A) Why don't you prepare light meals as well?
- B) Do you serve pizza with chicken and cheese?
- C) I prefer eating green vegetables, so can I order salad please?
- D) I should give up eating fried chicken and potato.
- E) Do you think a diet menu is better than starch food?

95. There is a beach near your home but the water is polluted and swimming is dangerous there. You see several people swimming, you say:

- A) If I were you, I would get permission before swimming in this water.
- B) Don't swim in the water. The water is very cold.
- C) Let's swim before the sun is up and it gets crowded.
- D) Before coming here, why didn't you take your swimming things?
- E) Don't swim in the water. The water is very dirty and you may get ill.

### **Appendix C. Interview (Turkish) Used in the Study**

1. Final sınavı test sorularının zorluk derecesi nasıldı?
2. Sınavın test bölümü dinleme becerileri, kelime bilgisi, dil bilgisi, okuduğunu anlama becerileri olmak üzere dört ana bölümden oluşmaktaydı.
  - a) Cevaplamakta en çok güçlük çektiğiniz bölüm hangisi? Neden?
  - b) Cevaplamakta kolaylık yaşadığınız bölüm hangisi? Neden?
3. Sınavın ayırt ediciliğiyle ilgili ne söyleyebilirsiniz?
  - a) Sizce sınav bilenle bilmeyeni, çalışanla çalışmayanı ayırt edici özelliğe sahip miydi?
4. Şıkların çeldiriciliği nasıldı?
  - a) Seçenekler arasında ikilemde kalma durumu yaşadınız mı? Neden?
5. Seçenekler arasında ikilemde kalınca ne hissettiniz?
  - a) İkilemde kaldığınız sorular sınavdaki diğer soruları yaparken motivasyonunuzu düşürdü mü?
6. Sınavla ilgili eklemek istediğiniz başka bir şey var mı?

## **Appendix D. Interview (English) Used in the Study**

- 1) How was the difficulty level of the multiple choice questions of the final exam?
- 2) In the multiple choice part of the exam, there were four main parts: listening skills, vocabulary, grammar and reading comprehension.
  - a) Which part was the most difficult for you? Why?
  - b) Which part was the easiest for you? Why?
- 3) What can you tell about the discriminating power of the exam?
  - a) Could the exam discriminate well between the good and poor or hardworking and lazy students?
- 4) How was the efficiency of the distractors?
  - a) Were you ever torn between two options (choices)? Why?
- 5) How did you feel when you were torn between two options?
  - a) Did the questions that you were torn between two options decrease your motivation while answering the other questions in the exam?
- 6) Is there anything else that you want to add about the exam?

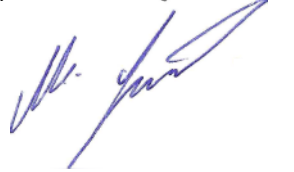
**T.C. SÜLEYMAN DEMİREL ÜNİVERSİTESİ**  
**EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE**

Burdur Mehmet Akif Ersoy Üniversitesi Yabancı Diller Yüksekokulu Müdürü olarak görev yaptığım 2015 yılı Bahar döneminde Öğretim Görevlisi Sibel TOKSÖZ yüksek lisans çalışmasında kullanmak üzere 2015 yılı Bahar döneminde hazırlık sınıflarına uygulanan Final sınavlarını incelemek istemiş ve tarafına sözlük olarak izin verilmiş, yazılı belge vermeye gerek duyulmamıştır.

Bilgilerinize sunar, gereğini arz ederim.

01/08/2018

Doç. Dr. Mustafa SEVIK



Burdur Mehmet Akif Ersoy Üniversitesi  
Eğitim Fakültesi / Yabancı Diller Eğitimi Bölümü  
Tel. 0 248 213 40 21  
Cep: 0506 351 37 73

## CURRICULUM VITAE

**Name Surname:** Sibel TOKSÖZ

**Birth Place and Date:** Kahramanmaraş, 1989

**Marital Status:** Married

**Foreign Language:** English

### **Education (Institution and Year)**

**Bachelor's Degree:** Boğaziçi University, English Language Education ( 2007-2011)

### **Conferences**

Ertunç, A & Bayır, S. (2013). *A Jar of Journal*. Süleyman Demirel University 'How to cook up delicious lessons: Some Ingredients'.

Karakaş, A., Toksöz, S., & Toksöz, İ. (2017). *Student English language teachers' research paper writing experiences*. Paper presented at the 2nd International Contemporary Education Research Congress. 28 September - 1 October 2017, Muğla Sıtkı Koçman University, Muğla, Turkey. (2017)

### **Publications**

Ertunç, A., Taş, B., Toksöz, S., & Zeybek, G. (2015). Never too late to mend: ELT teachers' thoughts on the teacher trainee curriculum. *Journal of Second and Multiple Language Acquisition-JSMULA* (ISSN: 2147-9747), 3(3).

Karakaş, A., Toksöz, S., & Toksöz, İ. (2017). Problems faced by pre-service English language teachers in writing research papers. *Balikesir University Journal of Social Sciences Institute*, 20(38), 369-385.

Toksöz, S. & Ertunç, A. (2017). Item Analysis of a Multiple-Choice Exam. *Advances in Language and Literary Studies*, 8(6), 141-146. (Indexed in ERIC)