

November 2019

Detection of Gender-Related Differential Item Functioning (DIF) in the Mathematics Subtests in Turkey

Zeynep Merve Sapmaz
University of South Florida

Follow this and additional works at: <https://digitalcommons.usf.edu/etd>



Part of the [Educational Assessment, Evaluation, and Research Commons](#)

Scholar Commons Citation

Sapmaz, Zeynep Merve, "Detection of Gender-Related Differential Item Functioning (DIF) in the Mathematics Subtests in Turkey" (2019). *USF Tampa Graduate Theses and Dissertations*.
<https://digitalcommons.usf.edu/etd/8099>

This Ed. Specialist is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

Detection of Gender-Related Differential Item Functioning (DIF) in the Mathematics Subtests in
Turkey

by

Zeynep Merve Sapmaz

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Education in Curriculum and Instruction
with a concentration in Measurement and Evaluation
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Robert F. Dedrick, Ph.D.
John Ferron, Ph.D.
Yi-Hsin Chen, Ph.D.

Date of Approval:
October 22, 2019

Keywords: differential item functioning, DIF, item bias, Cochran-Mantel-Haenszel, item response theory 2-PL IRT-likelihood ratio test, logistic regression.

Copyright © 2019, Zeynep Merve Sapmaz

Acknowledgments

I would especially like to thank the Republic of Turkey - Ministry of National Education for providing an opportunity for completing a master's degree in the USA and I would like to extend my appreciation to the ÖSYM (the MSPC) for providing the data of my thesis and supporting me throughout my study.

I would like to express my deepest gratitude to my major advisor, Dr. Robert F. Dedrick, for his patience, leadership, and invaluable suggestions.

I also would like to thank my committee member, Dr. Yi-Hsin Chen, for his unique guidance for the IRT analyses in my thesis, along with his patience and ideas. I thank my committee member, Dr. John Ferron, for his suggestions and supportive attitude. I also would like to thank my Turkish advisor, Dilara Bakan Kalaycıoğlu, for her attention, assistance, and suggestions.

In addition, I would like to thank our program specialist, Mr. Todd Williams, for his assistance.

I finally thank my father, Kemal Sapmaz, my mother, Ayse Sapmaz, and my siblings for their constant encouragement, love, and care.

Table of Contents

List of Tables	iii
List of Figures	vi
Abbreviations	vii
Abstract	viii
Chapter 1: Introduction	1
1.1. Background	1
1.2. Significance of the Study	4
1.3. Statement of the Purpose	4
1.4. Research Questions	5
1.4.1. Broad Research Questions	7
1.5. Definition of Terms	7
Chapter 2: Literature Review	9
2.1. Introduction	9
2.2. Theoretical Background of Differential Item Functioning Studies	11
2.2.1. Overviews of CTT and IRT	12
2.2.2. DIF Fundamentals	16
2.2.3. Item Response Functions in IRT	16
2.2.4. IRT Assumptions	17
2.2.5. IRT models	18
2.2.6. Estimation of Item and Population Parameters for Dichotomous Data	19
2.2.7. Ability (θ) Estimations	21
2.2.8. DIF Methods	21
2.3. Gender Differences in Mathematics Abilities	27
2.3.1. Previous DIF Studies in the World	29
2.3.2. Previous DIF Studies in Turkey	30
Chapter 3: Methods	33
3.1. Materials	33
3.2. Participants	35
3.3. Descriptive Statistics and Analysis of DIF	35
3.4. Cochran-Mantel-Haenszel Procedure (C-M-H)	36
3.5. Logistic Regression Procedure (LR)	38
3.6. 2-PL IRT-LR	40
3.7. Distractor and DIF Analysis	42
3.8. Differential Item Functioning	42

Chapter 4: Results	46
4.1. Fundamental Mathematics Subtest (FMS).....	46
4.1.1. Descriptive Analysis	46
4.1.2. Cochran-Mantel-Haenszel Procedure (C-M-H).....	48
4.1.3. Logistic Regression Procedure	51
4.1.4. 2-PL IRT-LR Procedure	55
4.2. Mathematics Subtest (MS).....	63
4.2.1. Descriptive Analysis	63
4.2.2. Cochran Mantel Haenszel Procedure (C-M-H)	65
4.2.3. Logistic Regression Procedure	68
4.2.4. 2-PL IRT-LR Procedure	71
Chapter 5: Discussion	80
5.1. Summary	80
5.2. Findings and Conclusions for Cochran-Mantel- Haenszel Analysis	81
5.2.1. Findings for Fundamental Mathematics and Mathematics Subtests based on C-M-H	82
5.3. Findings and Conclusions for Logistic Regression Analysis	83
5.3.1. Findings for Fundamental Mathematics and Mathematics Subtests based on LR.....	83
5.4. Conclusion for Fundamental Mathematics and Mathematics Subtests for Non-IRT Analysis	85
5.5. Findings and Conclusions for 2-PL IRT-LR Analysis	86
5.5.1. Findings for 2-PL IRT-LR analysis	88
5.5.2. Findings based on Two-Group Approach.....	89
5.5.3. Conclusions for 2-PL IRT-LR Analysis and Discussion between non-IRT and IRT Approaches	89
5.6. Recommendations for Future Research	92
References.....	93
Appendices.....	98
Appendix A: Item Characteristic Curves	99
Appendix B: Some Original and Translated Test Items in the MSPC- 2018 HEIE	101

List of Tables

Table 2.1.1. The Level of the Item Difficulty.....	13
Table 2.1.2. The Level of the Item Discrimination Coefficient (D-value).....	14
Table 2.1.3. Some DIF Methods based on Wiberg Classification.....	22
Table 2.1.4. An Example of a Contingency Table.....	23
Table 2.1.5. ETS Delta Scale for DIF Level.....	25
Table 3.1.1. Tests in HEIE and Numbers of Questions in Tests	34
Table 3.2.1. The population of Higher Education Institutions Examination in 2018.....	35
Table 3.4.1. SAS Output Delivery System (ODS) Table Names for C-M-H.....	37
Table 3.4.2. Classification of ETS Delta Scale Based on MH-DIF.....	38
Table 3.6.1. Item Response Models and Analysis for Dichotomous Data in the PROC IRT Procedure	40
Table 3.6.2. PROC IRT Features for the Constrained Baseline Method.....	41
Table 3.7.1. Research Questions and Statistical Analysis	42
Table 4.1.1. Frequency Distribution of Gender of Student for Fundamental Mathematics Subtest	47
Table 4.1.2. Descriptive Statistics for Fundamental Mathematics Subtest Items.....	47
Table 4.1.3. Results of Cochran-Mantel-Haenszel Analysis for Fundamental Mathematics Subtest Items	49
Table 4.1.4. The Items with DIF Categorization in the ETS Delta Scale.....	51
Table 4.1.5. Results of Logistic Regression Analysis for Fundamental Mathematics Subtest Items	52
Table 4.1.6. The Items with DIF Categorization in the ETS Delta Scale.....	54

Table 4.1.7. Comparison of Types of DIF based on Two Chi-square Methods	54
Table 4.1.8. Model Fit Statistics for FMS	55
Table 4.1.9. Item Parameter Estimate Ranges for Each Group	57
Table 4.1.10. Item Parameter Estimates for Each Gender	57
Table 4.1.11. Results of 2-PL IRT-LR Analysis for Fundamental Mathematics Subtest Items.....	59
Table 4.1.12. Comparison of Significant Differences between Manifest Groups on FMS Test Items Using 2-PL IRT-LR Model	61
Table 4.1.13. Comparison of Types of DIF based on Non-IRT and IRT-LR Methods.....	62
Table 4.1.14. The Conclusion of the Items, which are including DIF or not, based on the Two- Groups Approach.....	62
Table 4.2.1. Frequency Distribution of Gender of Student for Mathematics Subtest	63
Table 4.2.2. Descriptive Statistics for Mathematics Subtest Items.....	64
Table 4.2.3. Results of Cochran-Mantel-Haenszel Analysis for Mathematics Subtest Items	66
Table 4.2.4. The Items with DIF Categorization in the ETS Delta Scale.....	67
Table 4.2.5. Results of Logistic Regression Analysis for the Mathematics Subtest Items.....	68
Table 4.2.6. The Items with DIF Categorization in the ETS Delta Scale.....	70
Table 4.2.7. Comparison of Types of DIF based on Two Chi-square Methods.....	71
Table 4.2.8. Model Fit Statistics for MS.....	71
Table 4.2.9. Item Parameter Estimate Ranges for Each Group	72
Table 4.2.10. Item Parameter Estimate for Each Group.....	73
Table 4.2.11. Results of 2-PL IRT-LR Analysis for Mathematics Subtest Item	75
Table 4.2.12. Comparison of Significant Differences between Manifest Groups on MS Items Using 2-PL IRT-LR Model	77
Table 4.2.13. Comparison of Types of DIF based on Non-IRT and IRT-LR Methods.....	78

Table 4.2.14. The Conclusion of the Items, which are including DIF or not, based on the Two Groups Approach.....	78
Table 5.1.1. General Mathematics Subtopics	81
Table 5.5.1. All methods` Comparisons based on Subtopics of Items, which favor males or females.....	90
Table B.1. The FMS Items, which is Identified with DIF in All Methods.....	101
Table B.2. The MS Items, which is Identified with DIF in All Methods.....	108

List of Figures

Figure 2.1.1. Test development process.....	10
Figure 2.1.2. Graphically displaying a) an unbiased item and b) a biased item..	11
Figure 2.1.3. Item Characteristic Curve.....	14
Figure 2.1.4. Graphically displaying uniform and non-uniform DIFs.....	16
Figure A.1. Item Characteristic Curves for Items with Non-Uniform DIF in the FMS.	99
Figure A.2. Item Characteristic Curves for Items with Non-Uniform DIF in the MS.....	100

Abbreviations

AERA: American Educational Research Association.

APA: American Psychological Association.

NCME: National Council on Measurement in Education.

DTF: Differential test functioning.

DIF: Differential item functioning.

MSPC (*Turkish name: ÖSYM*): Measurement, Selection, and Placement Center.

MONE: Ministry of National Education.

CTT: Classical test theory.

IRT: Item response theory.

2- PL IRT-LR: 2- parameter logistical item response theory- likelihood ratio test.

HEIE (*Turkish name: YKS*): Higher Education Institutions Examination.

FMS: Fundamental mathematics subtest (*Turkish name: 2018 TYT/ Temel Matematik Testi*)

MS: Mathematics subtest (*Turkish name: 2018 AYT/ Matematik Testi*).

C-M-H: Cochran-Mantel-Haenszel.

LR: Logistic regression.

SEM: Standard error of measurement.

IRF: Item response function.

ICC: Item characteristic curve.

MML: Marginal maximum likelihood.

CML: Conditional maximum likelihood.

JMLE: Joint maximum likelihood.

ETS: Educational Testing Service.

Standards: Standards for Educational, Psychological Testing.

Abstract

The main goal of this study was to investigate differential item functioning by gender in the Fundamental Mathematics (FMS) and Mathematics subtests (MS) of the MSPC-2018 Higher Education Institutions Examination. Each test consists of 40 items and for both subtests random samples of 10,000 students were received from the MSPC separately. To compare non-IRT (Classical Test Theory) and Item Response Theory (IRT) approaches, Cochran-Mantel-Haenszel (C-M-H), Logistic Regression (LR), and 2-PL IRT-LR statistics were used.

For the FMS, C-M-H, LR, and 2-PL IRT-LR procedures identified 18, 16, and 10 out of 40 items that had DIF, respectively. Based on the non-IRT approaches, the items, which favor females, divided into three mathematics subtopics, which are *number*, *algebra*, and *geometry*. There were only two items, which were item 5 and item 11 in the *number* subtopic, in Category C (large DIF) based on ETS delta scale. On the other hand, the items, which favored males, divided into three mathematics subtopics, which were *arithmetic*, *advanced math*, and *geometry*. There were only two items, which were item 18 and item 29 in *arithmetic* and *advanced math* subtopics, respectively, in Category C based on the ETS delta scale. Based on 2-PL IRT-LR results, the items, which favored males, divided into same subtopics with non-IRT approach results.

For the FMS, females tend to outperform males in four-operation skills, whereas males have higher performance on higher level mathematics (i.e., problem-solving, analytical thinking) and arithmetic skills than females.

For the MS, C-M-H, LR, and 2-PL IRT-LR procedures identified 22, 18, and 9 out of 40 items that had DIF, respectively. Based on the non-IRT approaches, the items, which favored females, divided into three mathematics subtopics, which were *number*, *arithmetic*, and *algebra*. There were no items, that favored females, identified in Category C. On the other hand, the items that favored males, divided into two mathematics subtopics, which were *advanced math* and *geometry*. There were only two items, which were item 22 and item 30 in the *advanced math* and *geometry* subtopics, respectively, in Category C based on the ETS delta scale. Based on 2-PL IRT-LR results, for the nine items with DIF, item 1 favored females, whereas the other items favored male students.

To compare groups based on total scores, the two-group approach was used for both tests. After analyzing the items, which were flagged as DIF, item 10 in the FMS was identified as moderately difficult and not discriminating well item and items 16, 30, 31, and 37 in the MS were identified as very difficult and not discriminating well items. Therefore, those items were not categorized items with DIF, and they require revisiting.

Chapter 1

Introduction

1.1. Background

When an education program is designed by educators, stakeholders, and politicians, the main goal is to reach the *highest program efficiency* via good program design, customized costs, and well-arranged measurement instruments (Royse et al., 2009). Well-developed measurement instruments have an essential place in educational and psychological programs because they help measure intended program efficiency via educational outcomes. In the recent version of the *Standards for Educational, Psychological Testing (Standards)*, a test was characterized as “a device or procedure in which a sample of an examinees behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (2014, p. 2.).

In the historical context of modern testing, Dubois (1970) posits that the history of contemporary testing started with the Chinese Civil Service Examinations (2200 B.C.E), and the evaluation of individual differences in the early 19th and 20th centuries by American and European psychologists. The assessment of achievement by old European schools and colleagues had significant impacts on the testing development process (cited in Bandalos, 2018). By the beginning of the 1800s, the first intelligence testing scales were developed and improved by Alfred Binet and Theodore Simon in 1905, 1908, and 1911, respectively—which were called the *Binet-Simon Scale* (Bandalos, 2018). Drawing on the Binet-Simon scale, systematic measurement and evaluation studies started in most countries, such as Turkey, in the early 20th

century. For the United States, the psychologist, Lewis Terman from Stanford University, provided a significant contribution to Binet's scale, which was named the *Stanford-Binet scale* (Bandalos, 2018).

The development and analysis of standardized tests and other tests have been guided by the *Standards*, which were developed by the *American Educational Research Association* (AERA), the *American Psychological Association* (APA), and the *National Council on Measurement in Education* (NCME). The *Standards* emphasize validity, reliability, and fairness, operations in the test development process, and in testing applications, such as test administrations (2014).

Standardized tests have been criticized for various reasons, including those that relate to validity, reliability, and fairness. Validity is a fundamental feature for a test because validity refers to “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (*Standards*, 2014, p. 11). Reliability is defined in broader terms as “the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported” (*Standards*, 2014, p. 33). It is important because if test scores are not reliable, the scores cannot be valid. Another test foundation is “fairness,” which is highlighted in the *Standards*. The *Standards* stress fairness as a significant validity issue and should be taken into account in all testing processes, such as test development or test score interpretation. Fairness refers to equality in testing and requires gaining more accurate results from measurements (2014).

Test fairness issues may originate from test *bias*, which can be defined as invalidity or systematic errors in the measurement of the test for group members (Camilli & Shepard, 1994, p. 4). The main point with bias is that systematic errors in measurement provide an unfair benefit to

one of the subgroups, and this situation creates an unequal opportunity for test-takers. To reduce measurement bias and make improvements in testing, *differential item functioning* and *differential test functioning* studies have been conducted for years. Differential test functioning (DTF) refers to test functioning differences between manifest groups, such as males and females. Differential item functioning (DIF) occurs “when equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership” (*Standards*, 2014, p. 51). DIF studies are critical to insuring quality of the tests because biased items in the test have an adverse effect on the validity of the tests.

In Turkey, nationwide standardized tests are made by the Measurement, Selection, and Placement Center (MSPC) since the 1970s in the name of the Ministry of National Education (MONE). In the 1970s, MSPC only served to conduct Higher Education Institution Entrance exams, but later on MSPC extended its service network by including different nationwide exams for different institutions, and every year, approximately 10 million candidates take these national exams in total (Özer, 2018). Özer highlights that there is no other exam center in the world that is not only responsible for conducting exams, but also providing selection and placement services after the exams (2018). That is the main reason why MSPC has no flexible time to do improvements in the system. Therefore, MSPC has three strategic goals to fix system problems, which are increasing accessibility, transparency and legal accountability, and monitoring and improvement (Özer, 2018). Those three strategic goals are designed to provide equal, fair, and better opportunities for everyone. For this goal, MSPC stresses the importance of transparency and legal accountability because national exams directly affect test takers’ future.

1.2. Significance of the Study

In this study, the validity issue related to the fairness of the Fundamental Mathematics and Mathematics subtests in the MSPC- 2018 Higher Education Institutions Examination was evaluated using Classical Test Theory (CTT) and Item Response Theory (IRT). Using some techniques from both CTT and IRT allows comparing both theories` applications to real data. Also, identifying biased items in the Fundamental Mathematics and the Mathematics subtests under the Higher Education Institutions Examination in 2018 helps to improve the tests` quality, increase legal accountability and transparency, and provide fairer tests in the foreseeable future.

1.3. Statement of the Purpose

In this study, the main purpose is to analyze the Fundamental Mathematics and the Mathematics subtests in the MSPC-2018 Higher Education Institutions Examination by conducting DIF analyses by gender and comparing non-IRT (CTT) and IRT approaches. For non-IRT procedures, Cochran-Mantel-Haenszel and Logistic Regression techniques are used, whereas, for IRT approaches, 2 PL IRT-LR model is used.

1.3.1. To investigate the direction of DIF for each test item in the Fundamental Mathematics subtest of the 2018 Higher Education Institutions Examination.

1.3.2. To investigate the direction of DIF for each test item in the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination.

1.3.3. To compare non-IRT and IRT approaches for each subtest items, used in the DIF analyses.

1.4. Research Questions

1. What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?
2. What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is characterized as having uniform and non-uniform gender DIF using the Logistic Regression method?
3. Do the Cochran-Mantel-Haenszel and Logistic Regression technique results for DIF match each other in the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?
4. Are the IRT assumptions met for the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination data?
5. How do the difficulty and discrimination parameter estimations compare between male and female students in the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?
6. What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?
7. What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination showed gender DIF using all three methods?

8. What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?
9. What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is defined as having uniform and non-uniform gender DIF using the Logistic Regression method?
10. Do the Cochran-Mantel-Haenszel and Logistic Regression technique results match each other in identifying gender DIF for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?
11. Are the IRT assumptions met for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination data?
12. How do the difficulty and discrimination parameter estimations compare between male and female students for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?
13. What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?
14. What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination showed DIF using all three methods?

1.4.1. Broad Research Questions

- 1.1. For each test, what percentage of the items show gender DIF?
- 1.2. To what extent is there agreement in the identification of gender DIF using these 3 methods, which are Cochran-Mantel-Haenszel, Logistic Regression, and 2-PL IRT-LR?
- 1.3. To what extent is there agreement in the identification of uniform and non-uniform DIF using these 3 methods?

1.5. Definition of Terms

Standardized Test: A test that is developed, administered, and scored using prespecified and uniform procedures (Popham, 1999).

Classical Test Theory (CTT): A theory that focuses on test scores in which the following equation is used to represent observed scores:

$$X \text{ (observed score)} = T \text{ (true score)} + E \text{ (error)}.$$

Item Response Theory (IRT): A theory that focuses on the relationship between performance and abilities to answer an item correctly (Hambleton & Jones, 1993).

Item Characteristic Curve: A graph with an S-shape that shows the properties of item difficulty and item discrimination, and pseudo-guessing index (for a 3-parameter IRT model).

Cochran-Mantel Haenszel (C-M-H): A non-IRT approach that is related to the dependency of two variables in a $2 \times 2 \times k$ contingency table (Kamata & Vaughn, 2004). C-M-H is designed to evaluate uniform DIF.

Logistic Regression: A non-IRT approach that is used for detecting DIF between manifest subgroups for dichotomous items (Kamata & Vaughn, 2004). In contrast to the C-M-H, the logistic regression model includes both main effects and interaction effects between groups and matching criterion. Logistic regression method tests both uniform DIF and non-uniform DIF.

IRT Likelihood Ratio Model (IRT-LR): An IRT approach to detect DIF based on Likelihood Ratio Tests.

Chapter 2

Literature Review

2.1. Introduction

Tests are crucial sources of information that help us understand individuals or groups and make various decisions about these individuals and groups' placement, selection, progress, and status in academic and non-academic areas (e.g., subjective well-being). There are many kinds of tests, such as those measuring intelligence, attitude, aptitude, or the ability of individuals and groups alike, that can be used for different purposes. Even if some of those tests have a small effect on individuals or groups, standardized tests are significant for students, educators, and other stakeholders because they help to shape an individual's future achievements.

Test development is “the process of producing a measure of some aspect of an individual's knowledge, skills, abilities, interests, attitudes, or other characteristics by developing questions or tasks and combining them to form a test, according to a specified plan” (*Standards*, 2014, p. 75). Figure 2.1.1 presents the test development process, which is created based on the *Standards*, and Hambleton and Jones's test development process.

1. Definition of the test purpose (s)
2. Definition of content and format specifications
3. Creating test blueprint
4. Composing test item pool
5. Field testing the items
6. Revising of the items
7. Preliminary test development
8. Pilot tests with representative samples (reliability, validity, utility, practicality)
9. Final test development
10. Analyzing how the test is functioning
11. Developing guidelines for administration, scoring, and interpreting the scores.

*According to Hambleton and Jones (1993), CTT and IRT show essential differences in steps 5, 7, and 10.

Figure 2.1.1. Test development process.

In the overall test development and usage context, the primary concerns of test stakeholders are high reliability, validity, and fairness in the tests. Fairness is a validity issue. According to the *Standards*, a fair test is characterized as a test that has no advantage or disadvantage for some individuals or subgroups due to characteristics of the tests (2014).

Within the test development steps, creating a test item pool (step 4) is a crucial step because each item affects directly the psychometric properties of the tests (Philip & Ojo, 2017). For nationwide standardized tests, test developers prefer multiple-choice items due to an efficiency in measuring cognitive skills. Even if multiple-choice items have many advantages in terms of checking psychometric properties, like other item types, multiple-choice items also may be a threat to the fairness of the tests.

According to Camilli and Shepard, bias in a test can be defined as “invalidity or systematic error in how a test measures for members of a particular group” (1994, p .8). In other words, bias in tests discriminates among members of a group of test-takers. Test-takers may be characterized by different variables, such as race, gender, ethnicity, language, age, or disability status. Figure 2.1.2 displays biased and unbiased items graphically (Mellenbergh, 1989).

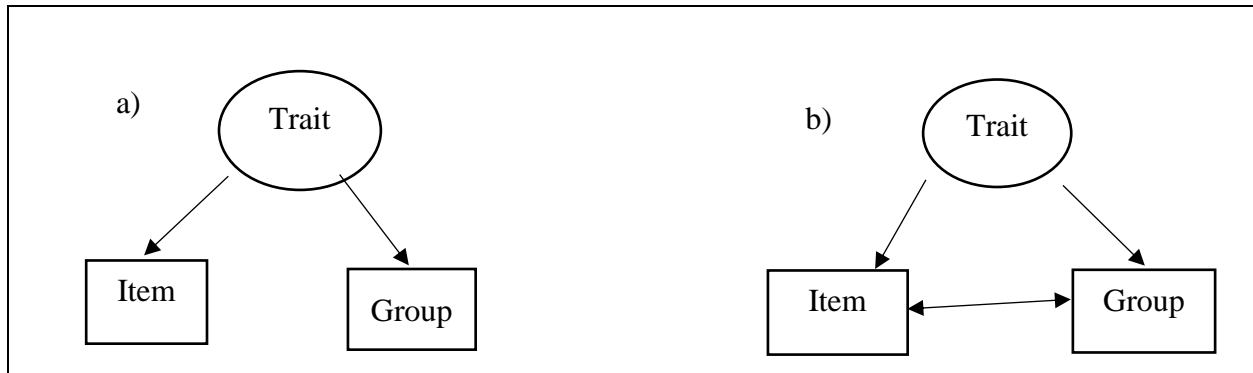


Figure 2.1.2. Graphically displaying a) an unbiased item and b) a biased item.

Mathematically, Mellenberg (1989) explains that if an item is biased,

$P(u=1|G, \theta) \neq P(u=1| \theta)$, and if an item is unbiased,

$P(u=1|G, \theta) = P(u=1| \theta)$.

Bias can be at the instrument (test)-level or item-level (De Ayala, 2008). If bias is evaluated at the instrument (test)-level, it is called *differential test functioning*, whereas if bias is evaluated at the item-level, it is called *differential item functioning (DIF)*. In other words, De Ayala defines DIF as “the method to detect items that are functioning differently across manifest groups of individuals” (2008, p. 324). Holland and Wainer have also defined DIF as occurring when “an item displays different statistical properties in different group settings” (1993, p. 4). DIF studies are placed in step 10, which is *analyzing how the test is functioning*, in the test development process.

2.2. Theoretical Background of Differential Item Functioning Studies

There are two psychometric theories, which are classical test theory and item response theory, currently used for addressing differential item functioning studies. Primarily, DIF studies were based on classical test theory applications, such as the ANOVA, delta-plot, transformed item difficulty, the Golden Rule procedures, etc. (Camilli & Shepard, 1994). However, these procedures are not currently recommended because, in classical test theory, item difficulty and

item discrimination indices are sample dependent (Camilli & Shepard, 1994). This situation leads to a change in an individual's performance based on the test difficulty. As an alternative to classical test theory, item response theory was mentioned by F.M. Lord in his dissertation in 1952 (Holland & Wainer, 1993, p. 8).

2.2.1. Overviews of CTT and IRT

According to Lord (1953), *observed* and *true* scores do not have the same meaning as *ability* scores. Ability scores are more essential than observed and true scores because observed and true scores are test-dependent, whereas ability scores are test-independent (as cited in Hambleton & Jones, 1993). In classical test theory, examinees' abilities on a test are based on observed (test) and true scores by using a simple linear equation to gain observed scores, which is X (test score) = T (true score) + E (error score). In this linear equation, true (T) and error (E) scores are identified as latent (unobserved) variables. The true score represents the score, which is free of all measurement errors. Because of the impossibility of this situation in measurement, the correct score is hypothetical or a latent rather than an observed score. On the other hand, the main challenge in CTT is measurement error (E) (Philip & Ojo, 2017). Measurement errors can be defined as "inconsistencies across test items, occasions, and raters" and CTT is used to describe the effects of measurement error on test scores (Bandalos, 2018, p. 158). Measurement errors affect test scores directly with an individual's true score, and some assumptions are required for addressing measurement error problems in CTT, which are:

1. The correlation between true (T) and error (E) scores is 0.
2. The mean error score for the population of examinees is 0.
3. The correlation between error scores on parallel tests is equal to 0.

In contrast with CTT, item response theory considers *ability* scores. Expressed another way, item response theory is interested in an individual's ability to answer an item, and abilities can remain at the same level for different tests unless being comprised of different conditions. Item response theory focuses on how performance related to the abilities is measured by the items in the test (Hambleton & Jones, 1993). Similarly, IRT ensures *the index of the precision of the test score*, which is *the standard error of measurement*, for everyone (DeMars, 2010).

2.2.1.1. Differences and Similarities between CTT and IRT

Item difficulty. Both CTT and IRT define item difficulty as the *probability of correct response* (DeMars, 2010). In CTT, the item difficulty range is 0 to 1.0, and it can be found as p , where:

$$p = \text{proportion of the people who responded to an item correctly.}$$

Table 2.1.1 presents the level of the item difficulty (p -value). If an item is too difficult or too easy, items may be revisited.

Table 2.1.1. *The Level of the Item Difficulty*

.80 and above	Easy item
.80- .30	Moderate item
.30 and below	Difficult item

In IRT, item difficulty is presented by the item parameter b . Figure 2.1.3, presents the location of parameter b in the item characteristic curve (ICC).

Item discrimination. In CTT, the item discrimination range is -1 to +1. In IRT, item discrimination is presented by the item parameter a . Figure 2.1.3, presents the location of the

parameter in the ICC. In IRT, item discrimination is called a *slope* (in the SAS output). Higher discrimination values show greater discrimination in both theory applications.

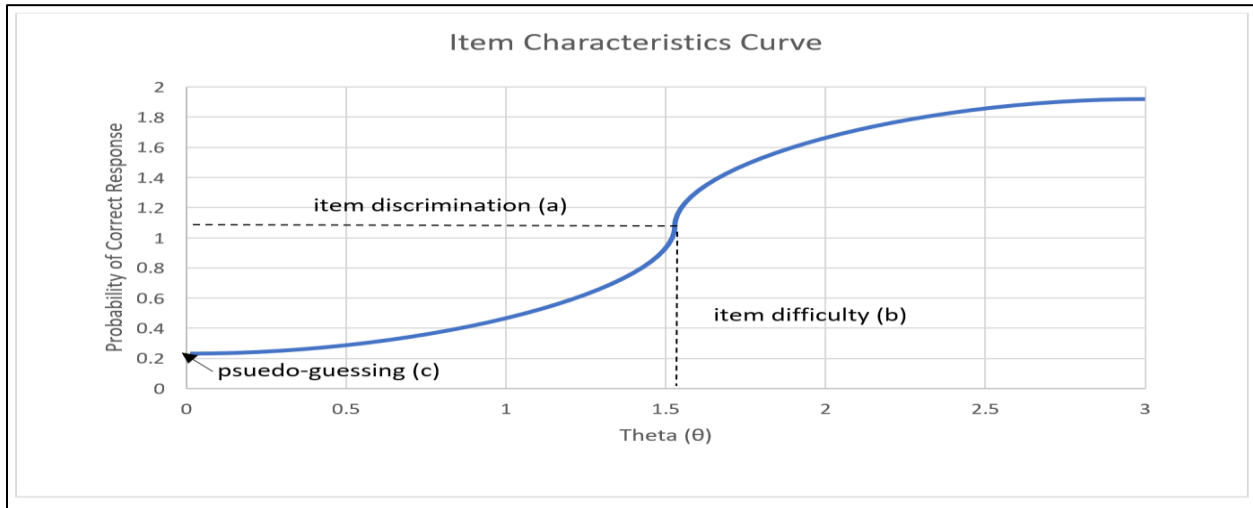


Figure 2.1.3. Item Characteristic Curve.

Computing item discrimination coefficients (D-value) helps to identify items with DIF, or items with poor construction (need to revisit). It means, if the item has weak or negative discrimination, it needs to be revisited, whereas if the item has good discrimination, it may be with DIF. For DIF analysis, to calculate the D-value the two ability (upper-lower) groups approach can be used (Chen et al., 2014). The formula of D-value is the percentage correct of the upper group - the percentage correct of the lower group. Table 2.1.2 presents the level of the item discrimination coefficient (D-value).

Table 2.1.2. *The Level of the Item Discrimination Coefficient (D-value)*

.30 and above	High discrimination
.0- .30	Moderate (little or no) discrimination
.0 and below	Negative discrimination

Reliability. In the overview of CTT and IRT, the standard error of measurement (SEM) is mentioned as the main challenge of CTT because of test dependent scores. In this section, the reliability of the test is explained with SEM for both theories.

In the CTT, reliability can be defined as a ratio of the true score variance to total (observed) score variance;

$$\frac{\sigma^2 T}{\sigma^2 T + \sigma^2 E}$$

Where:

$T = \text{True score}$

$E = \text{Error}$

Also, the SEM formula is:

$$SEM = SD * \sqrt{1 - Reliability}$$

According to DeMars, the standard error of measurement and reliability can be calculated with the information function in CTT and IRT. The higher the information function, the higher the reliability, whereas the higher the information function, the lower the standard error. At this point, IRT has an advantage because the information function can be calculated at the item-level (2010).

Parameter Invariance. Item response theory has a test-independence score; therefore, item parameters in different examinee populations should be the same (DeMars, 2010). There are several advantages of parameter invariance in IRT model parameters, which include being able to use these parameters in adaptive computer-based testing, comparing test-takers even if they are answering different items, and connecting different scales, which measure the same constructs (DeMars, 2010).

2.2.2. DIF Fundamentals

DIF occurs when manifest groups have a different “probability of answering correctly, although the group members have the same ability in the test” (Bandalos, 2018). In the DIF literature, the manifest groups are divided into *focal* and *reference groups*. In DIF studies, the focal group is usually identified as the minority or disadvantaged group, whereas the reference group is usually the majority or normative groups (Martinková et al., 2017). For instance, if a gender-related DIF study focuses bias against females, the reference group should be males and focal group should be females.

There are two types of DIF, which are uniform DIF and non-uniform DIF. Uniform DIF posits that the property is being measured consistently, whereas, non-uniform DIF stipulates that the property is being measured inconsistently (across). Figure 2.1.4. shows both uniform and non-uniform DIF.

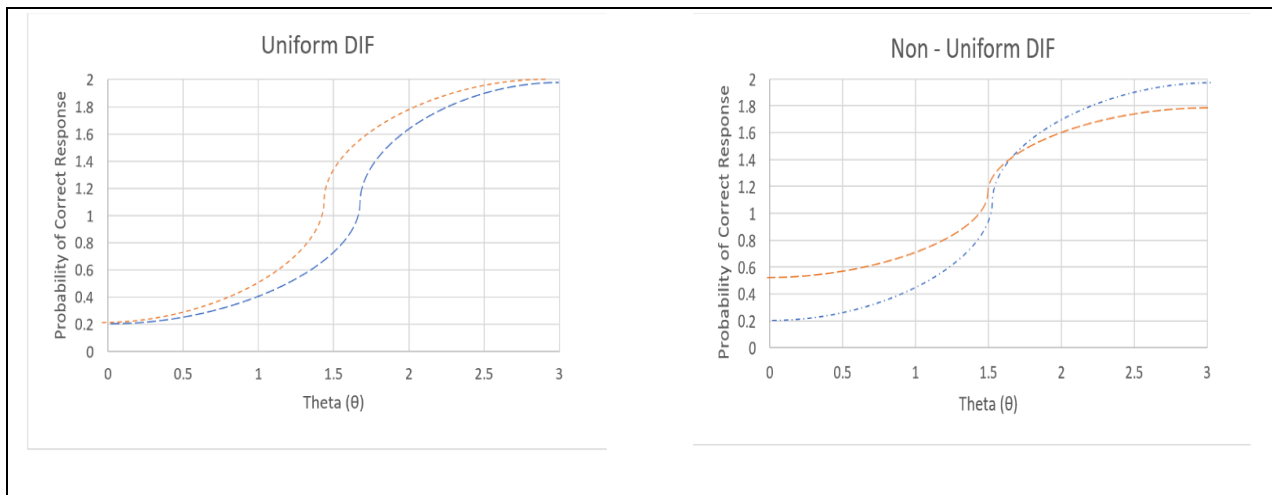


Figure 2.1.4. Graphically displaying uniform and non-uniform DIFs.

2.2.3. Item Response Functions in IRT

The item characteristic curve (ICC), also known as item response functions (IRF) presents the relationship between an individual’s ability (θ) and the probability of correct

response $P(\theta)$, which has an S-shape (Camilli & Shepard, 1994). The ICC can be defined by a four parameters logistic model, which includes item difficulty (b), item discrimination (a), pseudo-guessing (c), and ceiling (d) parameters.

2.2.4. IRT Assumptions

Item response theory represents a collection of mathematical models that indicate the relationship between item characteristics and individual abilities to the probability of a correct response to an item (Hambleton & Jones, 1993). Item response models can be used for dichotomous data or polytomous scored items and can be used for unidimensional or multidimensional data. The most common models in IRT are used for dichotomous items, which are the one-/Rasch, two-, three-, four- logistic parameter models. To choose which model is applied for the data, there are some assumptions required in IRT, which are *unidimensionality*, *local independence*, and *model specification*.

2.2.4.1. Unidimensionality

IRT models can be separated as unidimensional or multidimensional models. Accordingly, before choosing a model to analyze the data, it needs to take into account whether the model is eligible for the data and is aligned with the data set. Therefore, the first assumption requires to check unidimensionality, which means that “the model is characterized with a single parameter for each examinee, and other factors, which are affecting item responses, are not accepted and shared by other items” (DeMars, 2010, p. 38). According to DeMars, there are some techniques that may help to decide unidimensionality in the IRT, such as the analysis of the eigenvalues, Stout’s test of essential unidimensionality, etc. (2010). For the analysis of

eigenvalues, the inter-item correlation matrix, or polychoric correlation matrix (for SAS 9.4 program) can be considered.

2.2.4.2. Local /Conditional independence

In the SAS\STAT 14.3[®] User`s Guide book, local independency of the data can be evaluated by examining independency of observed responses (p. 4828). Basically, this assumption requires that after controlling the latent trait, item responses should be uncorrelated.

2.2.4.3. Model Specification

The final assumption is to identify the best model for the data. For dichotomous data, Rasch, one-, two-, three-, and four-parameter logistic models can be used to detect DIF.

2.2.5. IRT models

2.2.5.1. One-parameter model

One-parameter (1-PL) or Rasch model contains only the b parameter, which refers to item difficulty. In the 1-PL model, all item discrimination must be equal. The range of ability (θ) is generally between -3 to +3.

$$P(X = 1|\theta, \beta) = \frac{\exp(\theta - \beta)}{1 + \exp(\theta - \beta)}$$

i.e., Logit = Log p/(1-p) = Person Ability – Item Difficulty = $\theta - \beta$.

2.2.5.2. Two-parameter model

Two-parameter (2-PL) logistic model contains a and b parameters, which refer to item discrimination and item difficulty, respectively. The range of discrimination is generally 0 to 2 for multiple-choice items.

$$P (X = 1|\theta, \beta) = \frac{\exp D a(\theta- \beta)}{1 + \exp D a(\theta - \beta)}$$

i.e., D =scaling factor, which is 1.7 (it is used to make the logistic function close to the normal ogive function).

2.2.5.3. Three-parameter model

Three-parameter (3-PL) logistic model contains a, b, and c parameters, which refer to item discrimination, item difficulty, and pseudo-guessing, respectively. The range of the pseudo-guessing parameter is generally between 0 and 0.30.

$$P (X = 1|\theta, \beta) = c + (1 - c) \frac{\exp (\theta- \beta)}{1 + \exp (\theta - \beta)}$$

2.2.6. Estimation of Item and Population Parameters for Dichotomous Data

Item response theory explains N (*number of examinees*) $\times n$ (*number of items*) matrices by ability (θ) and item (β) parameters (Hambleton & Swaminathan, 1985). Before using dichotomous scored items for analysis, the items must be adjusted (item calibration) first, and then their parameters must be estimated. In this case, the estimation of scores uses likelihood functions, which are shown in the item characteristic curves, which are S-shaped (DeMars, 2010). For instance, for a correct response, the likelihood function is $P(\theta)$, and for an incorrect response, the likelihood function is $1 - P(\theta)$. Fundamentally, for estimation of ability (θ), Maximum Likelihood Estimation (ML) and Bayesian approaches are commonly used, and for item parameter estimations (β), Marginal Maximum Likelihood Estimation (MML), Conditional Maximum Likelihood (CML), and Joint Maximum Likelihood Estimation (JMLE) techniques are commonly used. One essential difference of Joint Maximum Likelihood Estimation (JMLE) from others is integrating person parameters into the likelihood function.

2.2.6.1. Marginal Maximum Likelihood

The marginal distribution in statistics is characterized as “the distribution of one variable after marginalizing over the distribution of another variable” (DeMars, 2010, p. 65). Stated another way, marginal maximum likelihood (MML) is the likelihood of the item parameters after marginalizing over ability (θ) (DeMars, 2010). In addition, MML gives full information about item response structures in the IRT.

2.2.6.2. Conditional Maximum Likelihood

Conditional maximum likelihood (CML) was developed in the context of the Rasch (1960) model and is used for different models, such as the Kelderman model (Cees & Glas, as cited in van der Linden, 2018). The critical difference between marginal maximum likelihood and conditional maximum likelihood is that CML is free of maximum-likelihood estimation assumptions, containing the distribution of the person parameters (Cees & Glas, as cited in van der Linden, 2018, p. 207).

2.2.6.3. Joint Maximum-Likelihood

Unlike marginal maximum likelihood parameter estimation, joint maximum likelihood estimation (JMLE) estimates both person ability (θ) and item parameters (De Ayala, 2008, p. 39). In other words, JMLE utilizes person estimates rather than marginalizing estimates (item parameter and ability (θ) estimation) (DeMars, 2010). JMLE is used commonly for the 1-PL or Rasch model.

2.2.7. Ability (θ) Estimations

2.2.7.1. Maximum Likelihood Estimation

To estimate likelihood functions for θ and β , maximum-likelihood estimation (ML) techniques are used, such as the Newton-Raphson algorithm (DeMars, 2010; Cees & Glas, as cited in van der Linden, 2018). By using ML techniques, two issues may occur, which are a large number of parameters and consistency of parameter estimations (Cees & Glas, as cited in van der Linden, 2018). In other words, both issues arise when the sample size increases, which results in increasing inconsistent item parameters evenly. Corresponding to this growing changeable item parameters, person parameters also increase, and it causes an issue for IRT applications (Cees & Glas, as cited in van der Linden, 2018).

2.2.7.2. Bayesian Approach

The *prior* distribution and *posterior* likelihood are fundamentally important in the Bayesian approach. Prior distribution refers to how ability is distributed in the population, whereas *posterior* likelihood has occurred if the *prior* distribution is multiplied by the likelihood function based on observed data (DeMars, 2010). The maximum estimate function or mean provides an estimate of ability. To estimate ability, if the maximum estimate function is utilized, it is called modal-a-posterior (MAP), whereas if a mean estimate is used, it is called expected-a-posterior (EAP) (DeMars, 2010).

2.2.8. DIF Methods

There are numerous DIF methods used for detecting differential item functioning. These methods can be separated as parametric or non-parametric, for observed or latent variables, usable for detecting uniform or non-uniform DIF, eligible for polytomous or dichotomous scored

data, and whether one can use a significance test of DIF or measure the size of DIF

(Test/Measure). Based on these features, Wiberg (2007) classified DIF methods and table 2.1.3

presented some of these DIF methods.

Table 2.1.3. *Some DIF Methods based on Wiberg Classification*

DIF Methods	Parametric/ Non- Parametric	Observed/ Latent variable	Dichotomous/ Polytomous	Test/ Measure	Uniform/ Non- Uniform
Mantel-Haenszel	Np	O	D/P	T/M	U
Standardization	Np	O	D	M	U
Chi-Square	Np	O	D	T	U
techniques					
SIBTEST	Np	L	D/P	T/M	U/N
Logistic	P	O	D/P	T/M	U/N
Regression					
Likelihood Ratio	P	O/L	D/P	T/M	U/N
Test					
General IRT-LR	P	L	D/P	T/M	U/N
IRT LRT	P	L	D/P	T	U/N
IRT methods	P	L	D/P	T/M	U/N
Lord`s Chi- squared test	P	L	D	T	U/N
Log-linear models	P	O	D/P	T	U/N
Mixed effect models	P	L	D/P	T	U/N

Note. P=Parametric; Np=Non-Parametric, O= Observed, L=Latent; D=Dichotomous, P=Polytomous; T=Test, M=Measure; U=Uniform, and N=Non-uniform.

In addition, in 1993, Wainer classified DIF methods as *Empirically Based* and *Model-Based Methods* (van der Linden, 2018). The Cochran-Mantel-Haenszel Procedure, Logistic Regression or Hierarchical Logistic Regression methods are the best known empirically based methods, whereas, IRT-LR methods, SIB Test, or Multilevel Bayesian IRT methods are the best known (IRT) model-based methods (Gamerman et al., as cited in van der Linden, 2018). So, in the present study, Mantel-Haenszel and Logistic Regression methods are considered as empirically-based non-IRT Methods for detecting DIF and IRT-LR are considered as (IRT) model-based methods.

2.2.8.1. Empirically Based Non-IRT DIF Methods

2.2.8.1.1. Cochran-Mantel-Haenszel Statistics (C-M-H)

The Cochran Mantel-Haenszel chi-square approach was investigated by Mantel and Haenszel as an alternative of matched-sample chi-square techniques (Mantel & Haenszel, 1959). This method has been adapted and improved for differential item functioning studies by Holland (1985) first and then by Holland and Thayer (1988) (Kamata & Vaughn, 2004). The C-M-H procedure is also named as a contingency table method ($2 \times 2 \times k$) and is a way of separating manifest groups (reference and focal group) based on the matching criterion (k), which is total test scores. In Table 2.1.4., an example of a contingency table for item t is given.

Table 2.1.4. *An Example of a Contingency Table*

Manifest Groups	Correct (1)	Incorrect (0)	Total
Reference Group	a_t	b_t	n_{rt}
Focal Group	c_t	d_t	n_{ft}
Total	n_{1t}	n_{0t}	N_t

In the M-H procedure, the null hypothesis (H_0) against the alternative (H_1) hypothesis:

$$H_1: \frac{Pr_t}{Qrt} = \alpha \frac{Pft}{Qft} \quad t=1,2, 3,\dots,k$$

DIF in the C-M-H procedure can be detected by these steps:

1. For $\alpha \neq 1$, and k is the number of levels of the matching criterion, the formula for estimating α is,

$$\hat{\alpha}_{m_H} = \frac{\Sigma (a_t d_t) / N_t}{\Sigma (b_t c_t) / N_t}$$

$\hat{\alpha}_{m_H}$ = common odds ratio.

According to Kamala and Vaughn (2004);

- ✓ If $\hat{\alpha}_{m_H}$ is equal to 1; it means, there is no difference between focal and reference groups based on the matching criterion level.
- ✓ If $\hat{\alpha}_{m_H}$ is higher than 1, it means “the indication of bias against the reference group.”
- ✓ If $\hat{\alpha}_{m_H}$ is lower than 1, it means “the indication of bias against the focal group.”

2. *Compute the signed index.*

In the second step, the common-odds ratio is converted to the signed index, which is natural log of the common odds ratio, and it is denoted by $\hat{\beta}_{MH}$.

The formula is $\hat{\beta}_{MH} = \ln(\hat{\alpha}_{m_H})$.

According to Kamala and Vaughn (2004);

- ✓ If $\hat{\alpha}_{m_H}$ is equal to 1, it means $\hat{\beta}_{MH} = 0$, there is no difference between focal and reference groups based on the matching criterion level.
- ✓ If $\hat{\alpha}_{m_H}$ is higher than 1, it means $\hat{\beta}_{MH}$ is a positive value (bias against the reference group).
- ✓ If $\hat{\alpha}_{m_H}$ is lower than 1, it means $\hat{\beta}_{MH}$ has a negative value (bias against the focal group).

3. *Convert a signed index to the magnitude of DIF.*

$$-2.35 \times \hat{\beta}_{MH} = MH-DIF (\hat{D})$$

4. *Determine DIF items using the Educational Testing Service delta metric scale.*

To evaluate the degree of DIF in the statistic, the ETS *delta metric* table can be used (Dorans & Holland, 1992).

Table 2.1.5. *ETS Delta Scale for DIF Level*

$ MH-DIF < 1$	Category A (negligible)
$1 < MH-DIF < 1.5$	Category B (moderate)
$ MH-DIF > 1.5$	Category C (large)

2.2.8.1.2. Logistic Regression Method

The logistic regression procedure was proposed by Rogers and Swaminathan (1990) to detect differential item functioning between manifest groups (reference and focal group). In this method, the outcome (dependent) variables can be identified as item responses (0 = incorrect, 1 = correct), whereas, the predictors (independent) can be defined as the total test score (matching criterion, k), manifest group membership (gender), and interaction between the total test and manifest group membership (Kamala & Vaughn, 2004). In the logistic regression procedure for DIF analysis, the predictors are added to models hierarchically. For instance, Model 1 and Model 2 (reduced models) represent main effects, which are total test scores and groups, respectively, whereas, Model 3 (full model) represents an interaction effect, which is an interaction between total test scores and groups. When interpreting results, the comparison of model 3 (full model) and model 2 (reduced model) should be checked first because of non-uniform DIF, and if the item doesn't include non-uniform DIF, the comparison of model 3 (full model) and model 1 (reduced model) should be checked because for uniform DIF. For checking uniform DIF, only model 2 may be used. If the uniform DIF also does not exist, the item can be identified as No-DIF. Model 3 can be shown as;

$$Y = \beta_0 + \beta_1 (\text{Ability}) + \beta_2 (\text{Gender}) + \beta_3 (\text{Ability} \times \text{Gender})$$

Model 2 is the reduced model, which includes two main effects:

$$Y = \beta_0 + \beta_1 (\text{Ability}) + \beta_2 (\text{Gender})$$

Model 1 is also a reduced model, which includes one main effect;

$$Y = \beta_0 + \beta_1 (\text{Ability})$$

2.2.8.2. Model-Based IRT DIF Methods

2.2.8.2.1. IRT-LR Method

The IRT Likelihood Ratio Test method has been proposed by Thissen et al. in 1988 and is available for both polytomous and dichotomous data to detect uniform and non-uniform DIF and DTF (Lopez, 2012). Fundamentally, in the IRT-LR, the null hypothesis is set up such that the item parameters between manifest groups do not differ and during the testing of the null hypothesis of no DIF, the compact and augmented models are compared (Thissen et al., as cited in Holland and Wainer, 1993). After that, the likelihood-ratio test statistic (G^2) is computed. If the p -value of G^2 is statistically significant, the item exhibits DIF.

In detail, W.-C Wang and Y. -L Yeh (2003), explain the application of compact and augmented models with three steps:

1. After providing the IRT model fit to the data, items (both anchor and studied) are constrained to have the same item parameters in both reference and focal groups (*compact model*). Then, the likelihood deviance of the Maximum Likelihood estimates is computed

$$(G^2_c = -2 \times \log\text{-likelihood}).$$

2. After providing the IRT model fit to the data, the items (both anchor and studied) are constrained to have the same item parameters in both reference and focal groups. However, there are no between-group equality constraints included in the item parameters (*augment model*). In other words, the augment model refers to allowing the item parameters to differ to best fit the

data for each group, after the IRT model is fit to the data separately for each group (Sireci & Rios, 2013). Then, the likelihood of deviance is computed (G^2_A).

3. The likelihood-ratio test statistic (G^2) is difference between the compact and augmented models, which means $G^2 = G^2_C - G^2_A$ (2003, p. 480).

In IRT-LR analysis, due to the sharply increasing number of anchor items, the power of DIF detection or Type 1 error rates can change. Therefore, Thissen et al. suggest two methods, which are constant anchor item method or free-baseline method and the all-other method or constrained baseline method, to gain high performance from the analysis (Lopez, 2012; W.-C Wang & Y. -L Yeh, 2003).

Firstly, the *constant anchor item method or free-baseline method* uses the anchor items that are kept constant throughout the item being studied (W.-C Wang & Y. -L Yeh, 2003). The method starts with a baseline model, which means the best model for fitting data (Lopez, 2012). Another method is known as the *all-other method or constrained baseline method*. To compare models for DIF analysis with this approach, the analysis starts with a baseline model that requires all item parameters constrained across manifest groups, and the models are created by releasing one item in sequence at a time (Lopez, 2012).

2.3. Gender Differences in Mathematics Abilities

In previous studies, gender differences in mathematics abilities were examined based on biological, cognitive, and psychosocial factors, such as individual experiences, socio-cultural or occupational factors (Geary, 1996; Geary, 1999; Halpern et al., 2007).

First of all, cognitive skills can be separated into *visuospatial*, *verbal*, and *quantitative* skills (Halpern et al., 2007). According to Halpern et al. (2007), visuospatial is a combination of visual and spatial skills, which include transforming, mental representing, mental rotating,

scanning pictures, etc. Verbal skills cover language usage, such as grammar, communication, comprehension, etc. (Halpern et al., 2007). Based on gender differences in cognitive skills, females have outperformed males in verbal skills (Halpern et al., 2007).

Second of all, some of the studies highlight that biological factors may affect a person's abilities in cognitive skills because of sex hormones (Baron-Cohen, 2005; Geary, 1999; Geary, 1996; Halpern et al., 2007). For instance, according to the Empathizing-Systemizing theory, the human brain can be formed as three types, which are *empathizing* (E), *sympathizing* (S) and *balanced* (B) brains (Baron-Cohen, 2005). By having particular brain types for each person, this theory supports that females may have an empathizing mind-type, and this type of brain comes with some advantages, such as driving to clarify someone's emotions, caring and treating other people, whereas, males may have a systemizing brain-type, which helps to analyze and operate a system (Baron-Cohen, 2005). Those advantages for males provide high abilities in the spatial and mathematical fields, while females outperform in their verbal skills (Baron-Cohen, 2005). However, according to Halpern et al. (2007), although androgen hormones provide advantages for males in the cognitive skills, males can be more able in mathematics than females, because of other reasons, such as individual interests, socioeconomic status, career choices, or cultural stereotypes. Selkowitz (1985) also found that personality variables may explain mathematical performance differences rather than biological sex differences because Selkowitz's findings imply that *masculine-oriented individuals* have higher mathematics performance than *female-oriented individuals*.

On the other hand, socio-cultural influences are also considered as a factor, which affects mathematics abilities in males and females. According to Geary (1996), biological differences (sex hormones) indirectly affect mathematical skills, but cultural stereotypes are directly guided

by gender interests. For instance, although the study found that there is no difference between female and male students in the elementary school in terms of cognitive abilities, because of stereotypes influences, females are less interested in mathematics course-taking and related activities in the following years (Geary, 1996).

In the previous studies, some researchers conclude that male examinees show higher performance for items that requires spatial skills than female examinees (Abedalaziz, 2010; Baran-Cohen, 2005; Geary, 1996; Halpern et al., 2007).

2.3.1. Previous DIF Studies in the World

Differential item functioning (DIF) studies have been extensive. There are some essential studies considered for this study, which are:

- ✓ In 2010, Abedalaziz used Logistic Regression and Mantel-Haenszel methods to investigate gender-related DIF in mathematics items. He concluded that males tend to show higher performance in spatial and deductive abilities, whereas females tend to show higher performance in numerical abilities.
- ✓ In 1997, Odett studied seventh-grade mathematics items (Michigan Educational Assessment Program “high stakes” test) using Mantel-Haenszel and 3-PL IRT approaches to investigate gender- and race-related DIF. For each technique, he used different mathematics items. As a result, when problem-solving and conceptualization abilities were required, males outperformed on fractions, percentages, and measurement subtopics. Also, females appeared to favor logical and statistical types of problems, if problem-solving or application abilities were required.

2.3.2. Previous DIF Studies in Turkey

The following studies focus on some MSPC exams with different levels, and these studies investigate gender-related DIF in mathematics subtests.

- In 2015, Yıldırım studied the 2012 year 8th Grade Level Determination Exam and investigated differential item functioning (DIF) based on gender and school types. For the DIF analysis, Cochran-Mantel-Haenszel and Logistic Regression methods were used. After that, to identify the significant level of DIF, and to reach a conclusion, the Delphi technique and item bias expert panel were used, respectively. Based on the gender-related DIF analysis for 20 Mathematics subtest items, one item (item 4) favoring girls was found and the following reasons were suggested:

- ✓ *The females enter the abstract stage earlier than the males,*
- ✓ *The conical shape, which is used in the item, is similar to the household items and the games girls play,*
- ✓ *The female students show higher performance for seeing details and overthinking than male students.*

On the other hand, one item (item 19) favored boys because

- ✓ *The item requires score calculation and is similar to the football score calculation system. Males are more interested in football games than females.*
- ✓ *The games, which are played by boys, improve their four-operation abilities in Mathematics.*

- In 2011, Kalaycıoğlu and Kelecioğlu studied the 2005 University Entrance Exam to detect gender-related DIF. For DIF analysis, Mantel-Haenszel and Logistic Regression

were used, and for the level of DIF, an expert panel method was used. According to the research results, Turkish subtest items have no DIF, whereas, social sciences subtest have seven items with DIF (one history and six philosophy), and mathematics and natural sciences have three items with DIF, respectively. One item from the natural sciences subtest (Physics item) was identified and favored male. The item includes automobile and speediness subtopics.

- In 2015, Şenferah researched the Mathematics Subtest of Level Determination Test in 2010 to investigate DIF analysis according to gender and school types. For DIF analyses, Mantel-Haenszel and Logistic Regression methods were used. After that, to reach a conclusion, the Delphi technique and item bias expert panel were used, respectively. According to MH and LR results, five items were identified as DIF and experts agreed that item 8 showed bias, which favored males because of some words, which are risk, factory, or occupational accident.
- Berberoğlu (1995) studied the Student, Selection, and Placement (SSP) exam mathematics subtest in 1992 based on gender and socio-cultural variables. The results showed that geometry items favored females, whereas, calculation and four-operation skills favored males.
- Yurdağül and Aşkar (2004) focused on the 2001 Secondary Schools Student Selection and Placement Examination subtests based on gender. Mantel-Haenszel was used, and they found 1 item with DIF in the Mathematics Subtest, which favored males. According to experts, this item is related to basketball, and it can be a potential source of bias.
- In 2011, Çepni investigated the Academic Staff and Postgraduate Education Entrance Examination Quantitative ability tests to conduct differential item functioning (DIF) and

differential bundle functioning (DBF) analysis. The Mantel Haenszel, logistic regression, SIBTEST, IRT-LR, and BILOG-MG DIF Algorithm methods were used. In conclusion, three items favored male students, whereas four items favored females in the Quantitative 1 Test. In the Quantitative 2 Test, one item revealed DIF, favoring males, whereas three items favored females. These results show that algorithmic operations, such as algebraic and abstract format, are more available for females, whereas the real-life problems are more available for males. Also, DBF analysis showed that four-operation items favored females, whereas word problems and the items, which required analytical thinking, favored males.

Chapter 3

Methods

This study was conducted to provide a comparison between some non-IRT and IRT DIF approaches and provide an evaluation of the two-parameter IRT logistic model using the Likelihood ratio test by the SAS 9.4 statistical software program for multiple-choice dichotomous test items. Non-IRT approaches, Cochran-Mantel-Haenszel and Logistic Regression, were used to detect differential item functioning (DIF). IRT approach, 2-PL's logistic IRT-LR method, was used to detect DIF.

3.1. Materials

The data were received as Microsoft Excel files in a CD from the Measurement, Selection, and Placement Center (MSPC) in Turkey. All statistical analyses, including descriptive statistics and DIF detecting analysis, were run with SAS 9.4 statistical software program.

The data used in this study were item responses from individuals tested on the MSPC-2018 Higher Education Institutions Examination (HEIE) that was developed for use in providing transmission to higher education for all candidates in Turkey. The MSPC-2018 HEIE in Turkey consists of three tests at different stages: The Basic Proficiency test, the Specialization Proficiency test, and the Foreign Language test. In table 3.1.1, details of all stages of the MSPC-2018 HEIE are represented (2018-HEIE Guide Book).

Table 3.1.1. *Tests in HEIE and Numbers of Questions in Tests*

	Sub-Tests	Number of Questions
Basic Proficiency Test (BPT) (<i>Turkish name: Temel Yeterlilik Testi</i>)	Turkish Language Test	40
	Social Sciences Test	20
	• History	5
	• Geography	5
	• Philosophy	5
	• Religious Culture and Moral Information (<i>or additional Philosophy questions</i>)	5
	Fundamental Mathematics Test	40
	Science Test	20
	• Physic	7
	• Chemical	7
• Biology	6	
Specialization Proficiency Test (SPT) (<i>Turkish name: Alan Yeterlilik Testi</i>)	Turkish Language Test- Social Sciences Test-1	40
	• Turkish Language and Literature	24
	• History-1	10
	• Geography-1	6
	Social Sciences Test- 2	40
	• History-2	11
	• Geography-2	11
	• Philosophy-2	12
	• Religious Culture and Moral Information (<i>or additional Philosophy questions</i>)	6
	Mathematics Test	40
Science Test	40	
• Physic	14	
• Chemical	13	
• Biology	13	
Foreign Language Test (FLT) (<i>Turkish name: Yabancı Dil Testi</i>)	Foreign Language Test	80

Note. Time for BFT, SPT, and FLT were limited by 135, 180, and 120 minutes, respectively.

In this study, the data were limited to the Fundamental Mathematics subtest in the BPT and the Mathematics subtest in the SPT under the MSPC - 2018 Higher Education Institutions Examination. Each test consists of 40 multiple-choice items with five alternatives.

3.2. Participants

Table 3.2.1 shows how many students applied and how many students' exams are considered valid based on the MSPC- 2018 Evaluation Report.

Table 3.2.1. *The population of Higher Education Institutions Examination in 2018*

Steps	The number of candidates who apply the exam	The number of candidates who are considered valid
Basic Proficiency Test (BPT)	2.381.412	2.260.273
Specialization Proficiency Test (SPT)	2.019.564	1.887.568
Foreign Language Test (FLT)	131.423	109.593

2018 HEIE (YKS) Evaluation Report.

A random sample of students taking the BPT and SPT exams was requested from the Measurement, Selection, and Placement Center (MSPC) database. A random sample of 10.000 students was received for the Fundamental Mathematics Subtest, and a random sample of 10.000 students was also collected for Mathematics Subtest. The samples, chosen for the differential item functioning studies, were not the same individuals. Data obtained for consideration were individual test item scores and gender.

3.3. Descriptive Statistics and Analysis of DIF

To provide preliminary information about the tests, descriptive statistics were calculated, including the mean, standard deviation, minimum and maximum, skewness and kurtosis values, and Cronbach`s alpha. Using the FREQ Procedure in SAS 9.4 program, item discrimination,

item difficulty, *p-value*, item characteristic curve, and missing values for each item were also calculated. In the study, item discrimination and item difficulty values were reported for each test. After the classification of test items, the tests' unidimensionality was evaluated. In addition, distractor analyses were conducted using the two-group approach.

In this study, three methods were used and compared to detect DIF. Two of the three methods, Cochran-Mantel-Haenszel and Logistic Regression, are non-IRT approaches and the last process, 2-PL IRT-LR, is an IRT approach.

DIF can be classified as uniform or non-uniform. The Cochran-Mantel-Haenszel method provides odds ratios, chi-square statistics, and is suitable for detecting uniform DIF. The Logistic regression method is also suitable to detecting non-uniform DIF. Both C-M-H and LR require that data include item responses (1=correct, 0 = incorrect), group membership (gender; male=1, female =2), and ability (total test score) variables. Additionally, to detect non-uniform DIF in the Logistic Regression method, an interaction variable is required, which is a combination of ability and group membership.

In conclusion, the Cochran-Mantel-Haenszel and the Logistic Regression methods were used, considering gender differences for comparison and confirmation of the two-parameter logistic model using SAS 9.4 statistical software program. In this study, manifest groups were identified as gender by assigning females to the focal group and males to the reference group.

3.4. Cochran-Mantel-Haenszel Procedure (C-M-H)

The Cochran-Mantel-Haenszel procedure (1959) was investigated by Holland and Thayer in 1988 as a technique for evaluating differential item functioning (Holland & Wainer,1993). The C-M-H method compares and matches manifest groups (focal and reference groups) based on a matching criterion, which is the total test score.

The FREQ procedure in SAS/STAT 13.1® was released in 2013. To create a table in the FREQ procedure, table names are referred to the Output Delivery System (ODS), and these table statements provide the contingency tables. For this study, output dataset table names and options are presented in table 3.4.1.

Table 3.4.1. SAS Output Delivery System (ODS) Table Names for C-M-H

Table Name	Description
C-M-H	Cochran-Mantel-Haenszel Statistics
BreslowDayTest (BDT)	Breslow-Day Test
CommonRelRisks	Common Relative Risks

(SAS/STAT 13.1® User Guide Book)

In the C-M-H procedure, chi-square (X^2) statistic and *common odds ratio*, α , (range is 0 to positive infinity) are provided, and the common odds ratio is the average of the number of possible test scores. The common odds ratio is usually transformed to the natural logarithm, β , (range is negative infinity to positive infinity) to place the value on a more interpretable scale. Proc FREQ is used to compute these indices (Penny). In this study, natural log odds ratios were calculated by the Microsoft Excel program. After transforming from common odds ratio to natural log odds ratio, the *delta scale*, which is determined by the Educational Testing Services (ETS), was used to investigate the level of DIF (Kamata & Vaughn, 2004). Delta scale formula is:

$$-2.35 \times \ln(\hat{\alpha}_{mH}) = \text{MH-DIF or } \hat{D}$$

Table 3.4.2 presents the classification of the ETS delta scale based on MH-DIF.

Table 3.4.2. *Classification of ETS Delta Scale Based on MH-DIF*

$ MH-DIF < 1$	Category A (negligible)
$1 < MH-DIF < 1.5$	Category B (moderate)
$ MH-DIF > 1.5$	Category C (large)

The Cochran-Mantel-Haenszel chi-square non-IRT test statistic results were compared to the more complex logistic regression and two-parameter logistic item response models using SAS 9.4 statistical software.

3.5. Logistic Regression Procedure (LR)

The logistic regression technique was proposed by Swaminathan and Rogers in 1990 (Gamerman et al. as cited in Van der Linden, 2018). The main differences with Cochran-Mantel-Haenszel are that logistic regression considers both uniform and non-uniform DIF and is more robust than C-M-H (Gamerman et al. as cited in Van der Linden, 2018). In the LR analysis, three models are computed and compared to investigate the existence of DIF.

Model 3 is the full model, which includes main effects and an interaction term;

$$Y = \beta_0 + \beta_1 (\text{Ability}) + \beta_2 (\text{Gender}) + \beta_3 (\text{Ability} \times \text{Gender})$$

Model 2 is the reduced model, which includes two main effects:

$$Y = \beta_0 + \beta_1 (\text{Ability}) + \beta_2 (\text{Gender})$$

Model 1 is also the reduced model, which includes one main effect:

$$Y = \beta_0 + \beta_1 (\text{Ability})$$

Comparing the full model (model 3) and the reduced model (model 2) is used to identify non-uniform DIF. If the item does not show non-uniform DIF, the full model (model 3) and

reduced model (model 1) should be compared to check uniform-DIF. If the uniform DIF does not exist, the item can be identified as No-DIF.

In the SAS 9.4 program, Logistic Regression was provided by the PROC LOGISTIC procedure. The model comparisons in the Logistic Regression can be evaluated using the Likelihood Ratio Test Chi-Squares (LRT- X^2) (Zhang, 2015).

In this study, model LRT- X^2 comparisons are made with the “ABS” function, and their p -values are found by the “PROBCHI” function in the SAS 9.4 program.

If the p -value for interaction, which is obtained by model 3 (LRT- X^2) - model 2 (LRT- X^2), is significant, it means that the item reveals non-uniform DIF.

If the p -value for main effects, which is obtained by model 3 (LRT- X^2) - model 1 (LRT- X^2), is significant, it means that the item reveals uniform DIF.

To determine which items favor girls or boys, the “Odds Ratio Estimates” table can be considered. Based on gender odds ratio values, the item can be identified as favoring males or females. In previous studies, focal and reference groups were coded with 0 and 1, respectively (Abedalaziz, 2010; Kamata & Vaughn, 2004). Therefore, C-M-H odds ratio was interpreted that if the significant odds ratio is higher than 1, the item shows DIF in favor of males, whereas, the item shows DIF in favor of females (focal groups = female, reference groups =male). However, in this study, focal and references groups were coded with 2 and 1, respectively, because IRT approach requires coding between 1 and 9999 for groups in SAS 9.4 program. Because of conducting all methods together, for the C-M-H and LR, if the significant odds ratio is higher than 1, the item shows DIF in favor of females, whereas, the item shows DIF in favor of males (focal groups = female, reference groups =male) in this study.

Like C-M-H analysis interpretation, the ETS delta scale can be considered for evaluating the DIF effect size.

3.6. 2-PL IRT-LR

The PROC IRT procedure in SAS/STAT 13.1 was released in 2013 to allow analyses of several item response models for both dichotomous and polytomous data. Choi presents the list of item response models in the PROC IRT procedure (2017). Based on Choi’s table, table 3.6.1 presents the item response models and analysis for dichotomous data in the PROC IRT procedure.

Table 3.6.1. *Item Response Models and Analysis for Dichotomous Data in the PROC IRT Procedure*

Model	Item		Parameters		Data
	Difficulty (Intercept)	Discrimination (Slope)	Pseudo- Guessing	Ceiling	Dichotomous
Rasch, 1- PL/PM	√				√
2-PL/PM	√	√			√
3-PL/PM	√	√	√		√
4-PL/PM	√	√	√	√	√
EFA/CFA for testing multidimensionality	√	√			√
Multigroup Analysis	√	√	√	√	√
Model fit					√
Item fit					√
					(Unidimensional only)

Note. IRT =Item response theory. PL/PM= Parameter logistic/probit model. EFA =Exploratory factor analysis. CFA= Confirmatory factor analysis.

In the PROC IRT procedure, multiple-group analysis can be performed with the BY or GROUP statements. These statements are used for separating sets of results for each group. For

this study, BETWEEN-GP was used in the EQUALITY statement to specify the subset of the groups in the multiple-group analysis.

In this study, an IRT based method, 2-PL IRT-LR was conducted, and for conducting a 2-PL IRT-LR test for DIF, a *constrained baseline* method was used. To implement the *constrained baseline* method, some specifications are defined in the PROC IRT procedure. Table 3.6.2 presents these specifications for the data.

Table 3.6.2. *PROC IRT Features for the Constrained Baseline Method*

Model used for Dichotomous Data	Calibration	Output
2-PL IRT-LR	Link function: Probit	Model Fit: AIC, BIC, Log-Likelihood, LR Chi-Square, and LR Chi-Square DF
	Item calibration: MML	Eigenvalues of the Polychromic Correlation Matrix
	Optimization Technique: Quasi-Newton	Iteration History
	Maximization Method: Adaptive Gauss-Hermite Quadrature	Item Parameter Estimates

Note. 2-PL IRT-LR = 2- parameter item response theory likelihood ratio. MML= Marginal Maximum Likelihood. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion.

The IRT-LR method compares the likelihood ratios of models and detects DIF using the likelihood ratio by testing a null hypothesis based on the comparison of item parameters of manifest groups (Atalay Kabasakal et al., 2014).

In this study, likelihood ratio comparisons are made with the “ABS” function, and their *p*-values are calculated by the “PROBCHI” function in the SAS 9.4 program. If the *p*-value for ab-DIF is statistically significant (*p*-value < .001), the item reveals non-uniform DIF. If item has no

non-uniform DIF, uniform DIF should be checked secondly. So, if *p-value* for b-DIF is statistically significant ($p\text{-value} < .001$), the item reveals uniform DIF. If both *p-value* for b-DIF and *p-value* for ab-DIF are not statistically significant, item reveals no DIF.

To detect gender-related DIF, parameter “b” can be compared because of this parameter refers to item difficulty (Odett, 1997). If the difference between the b parameters for reference and focal groups is positive, the item favored the focal group. Otherwise, if the difference between the b parameters for the reference and focal groups is negative, the item favored the reference group (focal group = female, reference group = male).

3.7. Distractor and DIF Analysis

After completing item and DIF analyses, items with DIF are evaluated based on distractors. The problems may come from item construction (good or poor item). So, items with DIF may require revisiting these items. Therefore, to evaluate items with DIF, a two-group approach was used.

3.8. Differential Item Functioning

In Table 3.7.1., research questions and responses are presented to detect differential item functioning for both subtests.

Table 3.7.1. *Research Questions and Statistical Analysis*

Research Questions	Variables	Statistical Analysis
1. What percentage of the items on the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	Cochran-Mantel-Haenszel method was used to test the null hypothesis ($H_0: \alpha_{mh} = 1$) for detecting DIF between manifest groups.

Table 3.7.1. (Continued)

Research Questions	Variables	Statistical Analysis
2. What percentage of the items on the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is characterized as having uniform and non-uniform DIF using the Logistic Regression method?	<u>Reference Group</u> Males	Logistic Regression method was used to test the null hypothesis ($H_0: \alpha_{mh} = 1$) for detecting DIF between manifest groups.
	<u>Focal Groups</u> Females	
3. Do the Cochran-Mantel-Haenszel and Logistic Regression technique results for DIF match each other in the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination?	<u>Reference Group</u> Males	Logistic Regression and Cochran-Mantel-Haenszel method results are compared and evaluated based on similarities and differences. ETS delta scale was used to identify the level of DIF for the biased items.
	<u>Focal Groups</u> Females	
4. Are the IRT assumptions met for the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination data?	<u>Reference Group</u> Males	To find the best IRT- LR model fit the data, SAS 9.4 was used. Three IRT-LR model assumptions are checked.
	<u>Focal Groups</u> Females	
5. How do the difficulty, and discrimination parameter estimations compare between male and female students in the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination?	<u>Reference Group</u> Males	2-PL IRT-LR model and SAS 9.4 were used for estimating a and b parameters and detecting differences in manifest groups in item responses if disagreements occur.
	<u>Focal Groups</u> Females	
6. What percentage of the items on the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?	Item parameters: a and b	2-PL IRT-LR model item parameters were estimated using the Marginal Maximum Likelihood for detecting DIF between manifest groups.
	<u>Reference Group</u> Males	
	<u>Focal Groups</u> Females	

Table 3.7.1 (Continued)

Research Questions	Variables	Statistical Analysis
7. What percentage of the items on the Fundamental Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination showed gender DIF using all three methods?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	Cochran-Mantel-Haenszel, Logistic Regression, and 2-PL IRT-LR results are compared and assessed in terms of similarities and differences for subgroups.
8. What percentage of the items on the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	Cochran-Mantel-Haenszel method was used to test the null hypothesis ($H_0: \alpha_{mh} = 1$) for detecting DIF between manifest groups.
9. What percentage of the items on the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is defined as having uniform and non-uniform gender DIF using the Logistic Regression method?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	Logistic Regression method was used to test the null hypothesis ($H_0: \alpha_{mh} = 1$) for detecting DIF between manifest groups.
10. Do the Cochran-Mantel-Haenszel and Logistic Regression technique results match each other in the identifying gender DIF for the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	Logistic Regression and Cochran-Mantel-Haenszel method results are compared and evaluated based on similarities and differences. ETS delta scale was used to identify the level of DIF for the biased items.
11. Are the IRT assumptions meet for the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination data?	<u>Reference Group</u> Males <u>Focal Groups</u> Females	To find the best IRT- LR model fit the data, SAS 9.4 was used. Three IRT-LR model assumptions are checked.

Table 3.7.1 (Continued)

Research Questions	Variables	Statistical Analysis
12. How do the difficulty, and discrimination parameter estimations compare between male and female students for the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination?	<u>Reference Group</u> Males	2-PL IRT-LR model and SAS 9.4 were used for estimating a and b parameters and detecting differences in manifest groups in item responses if disagreements occur.
	<u>Focal Groups</u> Females	
13. What percentage of the items on the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?	<u>Reference Group</u> Males	2-PL IRT-LR model item parameters were estimated using Marginal Maximum Likelihood for detecting DIF between manifest groups.
	<u>Focal Groups</u> Females	
14. What percentage of the items on the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination showed DIF using all three methods?	<u>Reference Group</u> Males	Cochran-Mantel-Haenszel, Logistic Regression, and 2-PL IRT-LR results are compared and assessed in terms of similarities and differences for subgroups.
	<u>Focal Groups</u> Females	

Overall, the research questions for this study can be collected based on three broad questions, which are;

- 1.1. For each test, what percentage of the items show gender DIF?
- 1.2. To what extent is there agreement in the identification of gender DIF using these 3 methods, which are Cochran-Mantel-Haenszel, Logistic Regression, and 2-PL IRT-LR?
- 1.3. To what extent is there agreement in the identification of uniform and non-uniform DIF using these 3 methods?

Chapter 4

Results

This chapter presents the data analysis results. Data were students' responses to multiple-choice test items from the MSPC-2018 Higher Education Institutions Examination Mathematics subtests in BPT and SPT exams in Turkey. The data were received from the Measurement Selection and Placement Center. The results of the examinees' responses by group relative to gender were reported.

For the data analyses, descriptive statistics, the Cochran-Mantel-Haenszel chi-square model, the Logistic Regression model, and the 2-PL IRT-LR model were reported for Fundamental Mathematics and Mathematics subtests separately.

4.1. Fundamental Mathematics Subtest (FMS)

4.1.1 Descriptive Analysis

The first part includes frequency distributions for the random sample of examinees based on gender. Table 4.1.1 represents the frequency distribution for the Fundamental Mathematics Subtest. The sample of students for FMS was approximately evenly distributed with 5.250 (52.5%) male and 4.750 (47.5 %) female students. There was no missing data for gender identification.

Table 4.1.1. *Frequency Distribution of Gender of Student for Fundamental Mathematics Subtest*

Gender of Student	Number	Percent
Male	5250	52.5
Female	4750	47.5
Total	10000	100.0

In the second part, the mean score and the standard deviation were 7.31 (out of a total of 40) and 7.95, respectively. Skewness and kurtosis results show that the distribution was positively skewed and leptokurtic (Skewness = 1.70, Kurtosis= 2.72). The standard error of measurement was found as 1.95 and Cronbach`s alpha of the FMS was .94 for the total group.

Table 4.1.2 presents the item difficulty (p), the standard deviation of items, and item discriminations (r). The difficulty indices range from .604 to .033. The mean difficulty of the test was .356, which shows the FMS is highly difficult for examinees. Mean discrimination of the test was .554, which shows the FMS is moderately discriminating for examinees.

Table 4.1.2. *Descriptive Statistics for Fundamental Mathematics Subtest Items*

Item No.	Item difficulty (p)	SD	Item discrimination (r)
1.	.440	.496	.539
2.	.486	.500	.548
3.	.232	.422	.645
4.	.508	.500	.560
5.	.368	.482	.628
6.	.190	.392	.642
7.	.193	.394	.513
8.	.152	.359	.600
9.	.280	.449	.620
10.	.604	.489	.488
11.	.199	.399	.687
12.	.153	.360	.609
13.	.179	.384	.489
14.	.131	.338	.590
15.	.149	.356	.664
16.	.194	.396	.573
17.	.071	.258	.479
18.	.188	.391	.604
19.	.117	.321	.462

Table 4.1.2. (Continued)

Item No.	Item difficulty (p)	<i>SD</i>	Item discrimination (r)
20.	.265	.441	.612
21.	.192	.394	.647
22.	.187	.390	.577
23.	.261	.439	.559
24.	.101	.301	.609
25.	.161	.368	.348
26.	.079	.270	.492
27.	.170	.375	.550
28.	.068	.252	.385
29.	.084	.277	.510
30.	.121	.326	.663
31.	.076	.265	.577
32.	.177	.382	.556
33.	.039	.193	.420
34.	.101	.301	.644
35.	.107	.309	.657
36.	.055	.228	.549
37.	.054	.226	.398
38.	.104	.306	.661
39.	.039	.194	.424
40.	.033	.179	.401

$N= 10.000.$

4.1.2. Cochran-Mantel-Haenszel Procedure (C-M-H)

Research Question 1: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?

The first research question in the study is associated with the Cochran-Mantel-Haenszel method, which was conducted with the SAS 9.4 statistical software program. Although the C-M-H statistic has been used frequently in educational measurement, a significant limitation of C-M-H is that the method is not suitable for detecting non-uniform DIF (Zhang, 2015). Therefore, the purpose of using the Cochran-Mantel-Haenszel method in this study is to identify uniform DIF in the Fundamental Mathematics Subtest (FMS) items.

To implement the C-M-H method, the PROC FREQ procedure in SAS 9.4 was used. If C-M-H *p-value* is less than a significant level ($p < .001$), and Breslow-Day Test for Homogeneity of the Odds Ratios' *p-value* is higher or equal than a significant level ($p \geq .001$), the item is indicating uniform DIF. *Odds Ratio* section in the C-M-H output helps to identify which item shows DIF for which gender. If the significant odds ratio is higher than 1, the item shows DIF in favor of females, whereas, the item shows DIF in favor of males if the odds ratio is less than 1 (focal groups = female (coding with 2), reference groups = male (coding with 1)). Table 4.1.3 presents the results of the C-M-H procedure for the FMS items.

Table 4.1.3. *Results of Cochran-Mantel-Haenszel Analysis for Fundamental Mathematics Subtest Items*

Item no.	C-M-H <i>p-value</i>	C-M-H Odds ratio	Log Odds ratio	MH-DIF	Breslow-Day Test <i>p-value</i>	Breslow-Day Test χ^2	95% CI	Conclusion
1.	0.0176	0.8875	-0.1193	-	0.0879	13.77	0.80, 0.97	No DIF
2.	<.0001*	1.2404	0.2154	-0.50	0.9273**	3.10	1.11, 1.37	Uni. DIF
3.	<.0001*	0.6575	-0.4193	0.98	0.5814**	6.59	0.58, 0.74	Uni. DIF
4.	<.0001*	1.6937	0.5269	-1.23	0.5739**	6.65	1.52, 1.88	Uni. DIF
5.	<.0001*	2.2524	0.8119	-1.90	0.5921**	6.49	2.00, 2.53	Uni. DIF
6.	0.4611	0.9516	-0.0496	-	0.6985	5.54	0.83, 1.08	No DIF
7.	<.0001*	0.6649	-0.4081	0.95	0.9815**	1.98	0.59, 0.74	Uni. DIF
8.	0.0060	1.2109	0.1913	-	0.7487	5.08	1.05, 1.38	No DIF
9.	<.0001*	1.6594	0.5064	-1.19	0.3642**	8.74	1.47, 1.86	Uni. DIF
10.	<.0001*	0.5624	-0.5755	1.35	0.0135**	19.27	0.50, 0.62	Uni. DIF
11.	<.0001*	2.0089	0.6975	-1.63	0.2943**	9.59	1.77, 2.31	Uni. DIF
12.	0.2998	1.0755	0.0727	-	0.4891	7.44	0.93, 1.23	No DIF
13.	0.0005*	1.2321	0.2087	-0.49	0.7618**	4.95	1.09, 1.38	Uni. DIF
14.	0.1075	1.1255	0.1182	-	0.1155	12.89	0.97, 1.29	No DIF
15.	<.0001*	1.6068	0.4742	-1.11	0.1566**	11.87	1.38, 1.86	Uni. DIF
16.	<.0001*	1.3024	0.2642	-0.62	0.0030**	23.25	1.15, 1.46	Uni. DIF
17.	0.4323	1.0708	0.0684	-	0.1803	11.39	0.90, 1.27	No DIF
18.	<.0001*	0.4687	-0.7577	1.78	0.2577**	10.10	0.41, 0.53	Uni. DIF
19.	0.0134	0.8399	-0.1744	-	0.0019	24.42	0.73, 0.96	No DIF
20.	0.0293	0.8810	-0.1267	-	0.0800	14.06	0.73, 0.96	No DIF
21.	0.0409	0.8719	-0.1370	-	0.1123	12.98	0.76, 0.99	No DIF
22.	<.0001*	0.6791	-0.3869	0.90	0.4020**	8.32	0.59, 0.76	Uni. DIF

Table 4.1.3. (Continued)

Item no.	C-M-H <i>p</i> -value	C-M-H Odds ratio	Log Odds ratio	MH-DIF	Breslow-Day Test <i>p</i> -value	Breslow-Day Test χ^2	95% CI	Conclusion
23.	0.8153	1.0130	0.0129	-	0.8901	3.61	0.90, 1.12	No DIF
24.	<.0001*	0.6318	-0.4591	1.07	0.0589**	15.01	0.53, 0.74	Uni. DIF
25.	0.5984	1.0312	0.0307	-	0.5261	7.09	0.92, 1.15	No DIF
26.	<.0001*	0.6451	-0.4383	1.03	0.1931**	11.15	0.54, 0.76	Uni. DIF
27.	0.4437	1.0497	0.0485	-	0.8311	4.27	0.92, 1.18	No DIF
28.	0.0046	0.7819	-0.2460	-	0.2571	10.11	0.65, 0.92	No DIF
29.	<.0001*	0.5400	-0.6161	1.44	0.0048**	22.02	0.45, 0.64	Uni. DIF
30.	0.6528	0.9636	-0.0370	-	0.4151	8.18	0.82, 1.13	No DIF
31.	0.6627	0.9604	-0.0404	-	0.4328	8.006	0.80, 1.15	No DIF
32.	0.0006*	0.8037	-0.2185	0.51	0.1505**	12.01	0.70, 0.90	Uni. DIF
33.	0.1411	0.8457	-0.1675	-	0.3262	9.19	0.67, 1.05	No DIF
34.	<.0001*	1.4457	0.3685	-0.86	0.5471**	6.90	1.21, 1.71	Uni. DIF
35.	0.0806	0.8580	-0.1531	-	0.2034	10.96	0.72, 1.01	No DIF
36.	0.4885	0.9296	-0.073	-	0.7157	5.38	0.75, 1.14	No DIF
37.	0.2447	0.8938	-0.1122	-	0.6583	5.01	0.73, 1.08	No DIF
38.	0.0345	1.2042	0.1858	-	0.2186	10.71	1.01, 1.43	No DIF
39.	0.8209	1.0258	0.0254	-	0.5130	7.22	0.82, 1.27	No DIF
40.	0.1117	0.8237	-0.1939	-	0.4759	6.56	0.64, 1.04	No DIF

Note.

1. *p*-value <.001.
2. Uni. DIF=Uniform DIF.
3. DF for C-M-H is 1 and DF for the Breslow-Day test is 8.
4. If the C-M-H *p*-value is <.001*, and the Breslow-Day test *p*-value is $\geq .001^{**}$, the item reveals uniform DIF.

Research Question 1 Response: Based on the C-M-H results, items, 2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 16, 18, 22, 24, 26, 29, 32, and 34 show evidence of uniform DIF. Therefore, 45% of the 40 items are identified as exhibiting uniform DIF. Items 2, 4, 5, 9, 11, 13, 15, 16, and 34 favor female examinees, whereas items 3, 7, 10, 18, 22, 24, 26, 29, and 32 favor male examinees. To classify the DIF level for the items with DIF, the natural log odds ratio was calculated first (see Table 4.1.3), and then the ETS delta scale was used with the formula, which is $-2.35 \times \ln(\hat{\alpha}MH) = MH-DIF$. In table 4.1.4, the items with DIF are categorized based on the ETS Delta scale.

Table 4.1.4. *The Items with DIF Categorization in the ETS Delta Scale*

	Item numbers favoring female examinees	Item numbers favoring male examinees
Category A (negligible)	2, 13, 16, 34	3, 7, 22, 32
Category B (moderate)	4, 9, 15	10, 24, 26, 29
Category C (large)	5, 11	18

4.1.3. Logistic Regression Procedure

Research Question 2: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is characterized as having uniform and non-uniform gender DIF using the Logistic Regression method?

The second research question is associated with the logistic regression method. The logistic regression method is more robust than the Cochran-Mantel-Haenszel method and can detect both uniform and non-uniform DIF (Gamerman et al., as cited in van der Linden, 2018). Therefore, the purpose of using the logistic regression method in this study was to identify uniform and non-uniform DIF and compare the results with the C-M-H method results for the Fundamental Mathematics Subtest items.

To implement the logistic regression method, the PROC LOGISTIC procedure in SAS 9.4 program was used. The results are evaluated based on the Likelihood Ratio Test. To interpret logistic regression results, firstly, the *p-value* for interaction should be examined for non-uniform DIF. If it is not statistically significant ($p > .001$), the *p-value* for the main effect in the model with gender and total score should be checked for evidence of uniform DIF. If *p-values* for both interaction and the main effect are not statistically significant ($p > .001$), the conclusion is that there is no DIF in the item.

After identifying items with non-uniform and uniform DIF, like the C-M-H procedure, the *Odds Ratio* table helps to clarify which item reveals DIF for which gender. If the significant odds ratio is greater than one, the item reveals DIF in favor of females, otherwise, the item shows DIF in favor of males (focal group=females, reference group= males). Table 4.1.5 presents the results of the Logistic regression procedure for the FMS items.

Table 4.1.5. *Results of Logistic Regression Analysis for Fundamental Mathematics Subtest Items*

Item no.	Model 1 χ^2	Model 2 χ^2	Model 3 χ^2	p-value for the main effect	p-value for interaction	Odds Ratio for Gender	Log Odds Ratio	MH- DIF	Conclusion
1.	3800.035	3810.342	3800.264	0.89181	0.00150	0.85	-0.16	-	No DIF
2.	4378.304	4387.759	4393.629	0.00047*	0.01540**	1.16	0.14	-0.34	Uni. DIF
3.	4165.374	4216.726	4213.866	0.00000*	0.09081**	0.63	-0.46	1.08	Uni. DIF
4.	4935.290	5017.735	5046.270	0.00000*	0.00000***	1.61	0.47	-1.11	Non-uni. DIF
5.	5101.257	5258.843	5289.875	0.00000*	0.00000***	2.05	0.71	-1.68	Non-uni. DIF
6.	3780.959	3782.558	3781.529	0.75169	0.31049	0.91	-0.09	-	No DIF
7.	2290.713	2348.306	2330.165	0.00000*	0.00002***	0.63	-0.46	1.08	Non-uni. DIF
8.	2992.217	2997.698	2998.176	0.05083	0.48946	1.18	0.16	-	No DIF
9.	4122.213	4184.145	4187.959	0.00000*	0.05083**	1.58	0.45	-1.07	Uni. DIF
10.	4045.636	4168.942	4109.645	0.00000*	0.00000***	0.57	-0.56	1.32	Non-uni. DIF
11.	4575.548	4668.088	4678.127	0.00000*	0.00153**	2.00	0.69	-1.62	Uni. DIF
12.	3098.702	3099.015	3098.725	0.98881	0.59020	1.04	0.03	-	No DIF
13.	2030.477	2038.241	2041.082	0.00498	0.09189	1.18	0.16	-	No DIF
14.	2766.280	2767.911	2769.964	0.15845	0.15187	1.10	0.09	-	No DIF
15.	3763.439	3800.123	3791.187	0.00000*	0.00280**	1.61	0.47	-1.11	Uni. DIF
16.	2921.064	2935.503	2952.177	0.00000*	0.00004***	1.26	0.23	-0.54	Non-uni. DIF
17.	1573.901	1574.349	1579.652	0.05641	0.02130	1.06	0.05	-	No DIF
18.	3264.135	3417.098	3383.678	0.00000*	0.00000***	0.43	-0.84	1.98	Non-uni. DIF
19.	1623.184	1631.409	1625.005	0.40234	0.01139	0.81	-0.21	-	No DIF
20.	3880.473	3889.196	3887.390	0.03147	0.17899	0.84	-0.17	-	No DIF
21.	3869.165	3876.079	3876.990	0.02000	0.33988	0.83	-0.18	-	No DIF
22.	2936.604	2984.975	2964.212	0.00000*	0.00001***	0.63	-0.46	1.08	Non-uni. DIF

Table 4.1.5. (Continued)

Item no.	Model 1 χ^2	Model 2 χ^2	Model 3 χ^2	<i>p</i> -value for the main effect	<i>p</i> -value for interaction	Odds Ratio for Gender	Log Odds Ratio	MH- DIF	Conclusion
23.	3085.619	3086.186	3085.930	0.85578	0.61295	0.95	-0.05	-	No DIF
24.	2770.521	2806.426	2790.918	0.00004*	0.00008***	0.58	-0.54	1.28	Non-uni. DIF
25.	1001.551	1001.604	1003.197	0.43897	0.20680	0.98	-0.02	-	No DIF
26.	1697.350	1728.498	1722.875	0.00000*	0.01773**	0.60	-0.51	1.20	Uni. DIF
27.	2564.291	2564.300	2565.032	0.69049	0.39225	1.00	0	-	No DIF
28.	1031.016	1040.767	1033.024	0.36638	0.00539	0.75	-0.28	-	No DIF
29.	1838.145	1897.576	1867.980	0.00000*	0.00000***	0.50	-0.69	1.62	Non-uni. DIF
30.	3511.011	3511.669	3511.399	0.82373	0.60291	0.93	-0.07	-	No DIF
31.	2320.498	2320.775	2320.895	0.81967	0.72810	0.94	-0.06	-	No DIF
32.	2652.187	2667.449	2658.812	0.03642	0.00330	0.77	-0.26	-	No DIF
33.	1106.088	1108.437	1106.489	0.81822	0.16288	0.83	-0.18	-	No DIF
34.	3136.079	3153.969	3155.480	0.00006*	0.21908**	1.48	0.39	-0.92	Uni. DIF
35.	3334.001	3338.414	3339.416	0.06671	0.31682	0.82	-0.19	-	No DIF
36.	1978.293	1978.830	1978.399	0.94841	0.51124	0.91	-0.09	-	No DIF
37.	1052.895	1054.881	1054.646	0.41664	0.62724	0.86	-0.15	-	No DIF
38.	3352.751	3356.997	3356.893	0.12600	0.74773	1.21	0.19	-	No DIF
39.	1128.677	1128.772	1129.687	0.60351	0.33867	1.03	0.02	-	No DIF
40.	987.2139	990.0005	988.7028	0.47500	0.25463	0.80	-0.22	-	No DIF

Note. 1. df for *p*-value for the main effect is 2 and df for *p*-value for interaction is 1.

2. if *p*-value for main effect is $\geq .001$, the item reveals No DIF.

3. If *p*-value for main effect is $<.001^*$, and *p*-value for interaction is $>.001^{**}$, the item shows uniform DIF.

4. If *p*-value for the main effect is $<.001^*$, and *p*-value for interaction is $<.001^{***}$, the item reveals non-uniform DIF.

Research Question 2 Response: Based on the logistic regression results, items, 2, 3, 9, 11, 15, 26, and 34, indicate uniform-DIF. Items, 4, 5, 7, 10, 16, 18, 22, 24, and 29, indicate non-uniform-DIF. Therefore, 40% of the 40 items are identified as DIF. Items 2, 4, 5, 9, 11, 15, 16, and 34 favor female examinees, whereas the items 3, 7, 10, 18, 22, 24, 26, and 29 favor males.

To classify DIF level for the items with DIF, like the C-M-H method, the natural log odds ratio is calculated first, and then the ETS delta scale can be used with the formula, which is -2.35

$x \ln(\hat{\alpha}MH) = MH-DIF$. In table 4.1.6 (see Table 4.1.5), the items with DIF are categorized based on the ETS Delta scale.

Table 4.1.6. *The Items with DIF Categorization in the ETS Delta Scale*

	Item numbers favoring female examinees	Item numbers favoring male examinees
Category A (negligible)	2, 16, 34	-
Category B (moderate)	4, 9, 15	3, 7, 10, 22, 24, 26
Category C (large)	5, 11	18, 29

Research Question 3: Do the Cochran-Mantel-Haenszel and Logistic Regression technique results for DIF match each other in the Fundamental Mathematics Subtest of the MSPC-2018 Higher Education Institutions Examination?

Research Question 3 Response: C-M-H is not sensitive for detecting non-uniform DIF. Therefore, when comparing the two methods results, which items indicate DIF can be considered. Based on table 4.1.7, both methods detect DIF in the same items, except items 13 and 32.

Table 4.1.7. *Comparison of Types of DIF based on Two Chi-square Methods*

Methods	Items with Uniform DIF	Items with Non-Uniform DIF
<i>C-M-H</i>	2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 16, 18, 22, 24, 26, 29, 32, and 34.	-
<i>Logistic Regression</i>	2, 3, 9, 11, 15, 26, and 34.	4, 5, 7, 10, 16, 18, 22, 24, and 29.

Note. The bold items favor females.

4.1.4 2-PL IRT-LR Procedure

4.1.4.1 Checking Model Assumptions and Clarifying Which Model is Better for The Test

Research Question 4: Are the IRT assumptions met for the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination data?

There are three assumptions underlying IRT models. These assumptions are unidimensionality, local independence, and model-data-fit. Therefore, before analyzing the IRT-LR model, the assumptions were checked.

Research Question 4. Response: To evaluate the dimension of latent factors, *eigenvalues of the Polychoric Correlation Matrix* tables for each gender group were provided by the PROC IRT procedure. The tables show that there is only one dominant eigenvalue identified with 22.365 (the second eigenvalue is 2.73) for males and 23.477 (the second eigenvalue is 2.52) for females in the model, which supports model unidimensionality.

According to SAS/STAT 14® User Guide Book, independency of observed responses (items) is proof of the local independence assumption (p. 4828).

PROC IRT procedure supports response models for binary data, which are Rasch, one-, two-, three-, and four-parameter logistic models (Matlock Cole & Peak, 2017). Table 4.1.8 presents model fit statistics based on the models in the IRT-LR.

Table 4.1.8. *Model Fit Statistics for FMS*

	Rasch Model	1-PL	2-PL	3-PL	4-PL
<i>Log Likelihood</i>	-126343.5894	-126343.5883	-124639.9034	-124167.5859	-124617.262
<i>AIC</i>	252851.17871	252851.17659	249599.80687	248815.17182	248652.55778
<i>BIC</i>	253442.42662	253442.4245	250753.46133	250545.65351	250959.8667

Note. 1. *p*-value=.001.

2. AIC= Akaike`s information criterion (smaller is better).

3. BIC= Bayesian information criterion (smaller is better).

4. The bolded values are better.

To make a decision about which model is better fit, Log-Likelihood (LL), AIC, and BIC criteria were considered. In general, the standard tests with multiple choice items are more available for 2- or 3- PL IRT- LR model. Therefore, when comparing 2-, 3- and 4- PL IRT-LR models, the smaller log likelihood value is in 2-PL IRT- LR model. So, the 2-PL IRT-LR procedure was used for the FMS.

During the following IRT-LR procedures in the study, log-likelihood values for each parameter were compared to detect differential item functioning. Moreover, if the likelihood ratio chi-square and Pearson`s chi-square are included in the model fit table, it means, all response patterns are observed in the analysis (SAS/STAT 14.3 ® User Guide Book). For this study, all response patterns are not observed because Pearson`s chi-square statistic is not shown in the table.

Research Question 5: How do the difficulty, and discrimination parameter estimations compare between male and female students in the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Research Question 5 Response: PROC IRT provides the *Item Parameter Estimates* table, including difficulty and slopes estimates, standard errors, and *p*-values. While the difficulty parameter refers to item difficulty (b parameter), the slope parameter refers to item discrimination (a parameter). In table 4.1.9, a range of difficulty and discrimination parameter estimates were presented based on the groups. For the male examinees, most of the difficulty parameters are higher than 0, which suggests that most of the items in this test are relatively hard. Besides, for the female students, the difficulty parameters have higher estimates than males' difficulty parameter estimates. On the other hand, discrimination ranges for both groups suggest that all the items (responses) are adequate measures of latent traits.

Table 4.1.9. *Item Parameter Estimate Ranges for Each Group*

<i>Group</i>	Discrimination (a) Parameters Range	Difficulty (b) Parameters Range
Male	0.44 to 1.53	-0.62 to 2.77
Female	0.50 to 1.63	-0.12 to 2.97

In table 4.1.10, item discrimination (a) and item difficulty (b) parameter estimates are presented separately for both male and female examinees.

Table 4.1.10. *Item Parameter Estimates for Each Gender*

Item no.	b parameter for male	a parameter for male	b parameter for female	a parameter for female
1.	0.09	0.90	0.29	1.02
2.	-0.0008	1.07	0.05	1.13
3.	0.75	1.26	1.08	1.21
4.	-0.01	1.19	-0.08	1.30
5.	0.46	1.38	0.29	1.50
6.	1.01	1.28	1.16	1.26
7.	1.19	0.78	1.62	0.78
8.	1.33	1.14	1.37	1.13
9.	0.76	1.21	0.66	1.24
10.	-0.62	0.95	-0.12	1.13
11.	1.03	1.50	0.88	1.63
12.	1.26	1.19	1.41	1.12
13.	1.52	0.74	1.46	0.79
14.	1.48	1.08	1.52	1.12
15.	1.25	1.46	1.23	1.36
16.	1.27	0.89	1.15	1.03
17.	2.26	0.82	2.10	1.01
18.	0.94	1.10	1.46	1.09
19.	1.93	0.70	2.17	0.71
20.	0.71	1.12	0.92	1.10
21.	0.97	1.32	1.18	1.25
22.	1.08	0.98	1.40	1.02
23.	0.84	0.94	0.98	0.93
24.	1.51	1.18	1.81	1.22
25.	2.37	0.44	2.25	0.50
26.	1.93	0.90	2.43	0.83
27.	1.34	0.91	1.40	0.97
28.	2.69	0.61	2.89	0.65

Table 4.1.10 (Continued)

Item no.	b parameter for male	a parameter for male	b parameter for female	a parameter for female
29.	1.88	0.87	2.30	0.92
30.	1.31	1.49	1.47	1.44
31.	1.80	1.18	1.99	1.11
32.	1.27	0.86	1.49	0.90
33.	2.68	0.83	2.80	0.89
34.	1.56	1.38	1.51	1.47
35.	1.38	1.53	1.61	1.38
36.	2.10	1.09	2.13	1.26
37.	2.53	0.78	2.79	0.76
38.	1.46	1.47	1.50	1.53
39.	2.77	0.81	2.65	0.94
40.	2.74	0.86	2.97	0.86

Research Question 6: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?

Two-Parameter Logistic Analysis via IRT-LR Using SAS 9.4.

The research question 6 is associated with the two-parameter logistic model using the Likelihood Ratio test, and its ability to detect differences between groups while considering the examinee's ability, item discrimination and item difficulty parameters.

To implement the 2-PL IRT-LR method, the PROC IRT procedure in SAS 9.4 was used. The results are interpreted based on the Log-Likelihood (LL) values, which are general model fit LL, freely estimated intercepts' LL, and freely estimated intercept and slopes' LL (*constrained baseline method*). For both ab-DIF (non-uniform DIF) and b-DIF (uniform DIF), *p*-values were computed.

Table 4.1.11 presents the results of the 2-PL IRT-LR analysis of the FMS items. To conduct the analyses, *p*-value for ab-DIF should be looked first to determine statistical

significance ($p < .001$). If the p -value for ab-DIF is less than significant level ($p < .001$), the item including non-uniform DIF. If p -value for ab-DIF is not less than at significance level ($p < .001$) and if p -value for b-DIF is less than at significance level ($p < .001$), the item shows evidence of uniform DIF. If p -value for b-DIF is less than 0, and p -value for ab-DIF is $> .001$, the item reveals No DIF. The table 4.1.11 presents the results of the 2-PL IRT-LR analysis of the FMS items.

Table 4.1.11. Results of 2-PL IRT-LR Analysis for Fundamental Mathematics Subtest Items

Item no.	Intercept LL	Intercept and Slope LL	p -value for b-DIF	p -value for ab-DIF	Conclusion
1.	-124660.4825	-124662.8291	0.00004*	0.12556**	Uniform DIF
2.	-124641.5148	-124642.1535	0.52216	0.42419	No DIF
3.	-124665.6829	-124678.9745	0.00000*	0.00027***	Non-uniform DIF
4.	-124642.7915	-124644.2194	0.22930	0.23210	No DIF
5.	-124648.9254	-124653.6657	0.00325	0.02946	No DIF
6.	-124644.7612	-124648.0482	0.04311	0.06983	No DIF
7.	-124673.7347	-124681.5047	0.00000*	0.00531**	Uniform DIF
8.	-124640.0659	-124640.3159	0.93765	0.61708	No DIF
9.	-124641.9249	-124643.0737	0.36612	0.28380	No DIF
10.	-124711.8009	-124731.6418	0.00000*	0.00001***	Non-uniform DIF
11.	-124641.0694	-124646.5596	0.08370	0.01912	No DIF
12.	-124640.4777	-124643.4886	0.30988	0.08270	No DIF
13.	-124640.0331	-124640.4865	0.90029	0.50070	No DIF
14.	-124641.4716	-124641.6448	0.62778	0.67731	No DIF
15.	-124642.2334	-124642.2572	0.50230	0.87741	No DIF
16.	-124640.7501	-124643.7324	0.28054	0.08418	No DIF
17.	-124646.3759	-124646.6511	0.08039	0.59987	No DIF
18.	-124696.3212	-124722.9047	0.00000*	0.00000***	Non-uniform DIF
19.	-124647.8825	-124650.4414	0.01451	0.10968	No DIF
20.	-124652.8091	-124656.1016	0.00103	0.06960	No DIF
21.	-124645.9246	-124653.4521	0.00359	0.00608	No DIF
22.	-124670.2425	-124677.3816	0.00000*	0.00754**	Uniform DIF
23.	-124645.4747	-124646.557	0.08380	0.29819	No DIF
24.	-124654.8963	-124667.7358	0.00000*	0.00034***	Non-uniform DIF
25.	-124643.0423	-124643.4975	0.30876	0.49987	No DIF
26.	-124648.8816	-124663.7329	0.00003*	0.00012***	Non-uniform DIF

Table 4.1.11. (Continued)

Item no.	Intercept LL	Intercept and Slope LL	<i>p</i> -value for b- DIF	<i>p</i> -value for ab-DIF	Conclusion
27.	-124643.902	-124643.99	0.25226	0.76671	No DIF
28.	-124648.9317	-124651.1175	0.01062	0.13928	No DIF
29.	-124666.9584	-124681.5775	0.00000*	0.00013***	Non-uniform DIF
30.	-124641.578	-124646.2002	0.09803	0.03156	No DIF
31.	-124640.4376	-124643.8321	0.26927	0.06541	No DIF
32.	-124656.3375	-124658.725	0.00030*	0.12231**	Uniform DIF
33.	-124643.8977	-124645.7604	0.11878	0.17231	No DIF
34.	-124640.0729	-124640.4429	0.91013	0.54301	No DIF
35.	-124640.6567	-124649.7938	0.01952	0.00250	No DIF
36.	-124646.4036	-124647.3565	0.05878	0.32896	No DIF
37.	-124641.486	-124644.2931	0.22234	0.09385	No DIF
38.	-124641.0596	-124641.4366	0.67464	0.53923	No DIF
39.	-124643.2665	-124643.2873	0.33614	0.88534	No DIF
40.	-124641.9605	-124644.9629	0.16750	0.08314	No DIF

Note. 1. *p*-value = .001.

2. LL= Log likelihood.

3. General Log likelihood = -124639.9034.

4. if *p*-value for b-DIF is < 0, and *p*-value for *ab*-DIF is >.001, the item reveals No DIF.

5. If *p* value for b-DIF is <.001*, and *p*-values for *ab*-DIF is > .001**, the item reveals Uniform DIF.

6. If *p* value for b-DIF is <.001*, and *p*-value for *ab*-DIF is < .001***, the item reveals Non-Uniform DIF.

Research Question 6 Response: Based on 2-PL IRT-LR results, items, 1, 7, 22, and 32 indicate uniform-DIF. Also, items, 3, 10, 18, 24, 26, and 29 are flagged as non-uniform-DIF.

Therefore, 25 % of the 40 items are identified DIF.

Ten items were flagged for DIF in the FMS subtest. To check which item favors which gender, parameter b can be compared because parameter b refers to item difficulty (Odett, 1997).

If the difference between the b parameters for reference and focal groups is positive, it can be said that the item favored the focal group. Otherwise, if the difference between the b parameters for the reference and focal groups is negative, the item favored the reference group (focal group = female, reference group = male). Table 4.1.12 presents the comparison of significant differences between manifest groups on the FMS test items using the 2-PL IRT-LR model.

Table 4.1.12. Comparison of Significant Differences between Manifest Groups on FMS Test Items Using 2-PL IRT-LR Model

Test Items by DIF	Females Parameter “b”	Males Parameter “b”	Difference in the “b” parameter
1.	0.29	0.09	-0.20
3.	1.08	0.75	-0.33
7.	1.62	1.19	-0.43
10.	-0.12	-0.62	-0.50
18.	1.46	0.94	-0.52
22.	1.40	1.08	-0.32
24.	1.81	1.51	-0.30
26.	2.43	1.93	-0.50
29.	2.30	1.88	-0.42
32.	1.49	1.27	-0.22

$p < .001$.

According to Table 4.1.12, all items with DIF favored males because their b parameter differences are negative. However, to understand which item with non-uniform DIF favored which gender, it needs to check the items in the ability scale because non-uniform DIF posits that the property is being measured inconsistently. Therefore, items with non-uniform DIF, which are 3, 10, 18, 24, 26, and 29 were evaluated based on the ICC (see Appendix A, figure A.1.).

According to ICCs, the items favored high ability group, which is reference (male) group, except item 10.

Research Question 7: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC-2018 Higher Education Institutions Examination showed gender DIF using all three methods?

Research Question 7 Response: The final research question in the study is associated with comparing non-IRT, and IRT approaches result based on how many items reveal DIF in their results. In table 4.1.13, a comparison of the three methods is presented.

Table 4.1.13. Comparison of Types of DIF based on Non-IRT and IRT-LR Methods

Methods	Items with Uniform DIF	Items with Non-Uniform DIF	Percentage of DIF
<i>C-M-H</i>	2, 3, 4, 5, 7, 9, 10, 11, 13, 15, 16, 18, 22, 24, 26, 29, 32, and 34.	–	45%
<i>Logistic Regression</i>	2, 3, 9, 11, 15, 26, and 34.	4, 5, 7, 10, 16, 18, 22, 24, and 29.	40%
<i>2-PL IRT-LR</i>	1,7, 22, and 32.	3, 10, 18, 24, 26, and 29.	25%

$p < .001$.

After DIF analysis, the items with DIF need to be compared on their p -value and D -values for conclusion. Table 4.1.14 presents the conclusion of the items, which are including DIF or not, based on two-group approach.

Table 4.1.14. The Conclusion of the Items, which are including DIF or not, based on the Two-Groups Approach.

Item no.	p -value	Num. Lower	Per. Lower	Num. Upper	Per. Upper	D-value
1.	.440	365	20.90%	2214	85.81%	.649
2.	.486	339	23.29%	2351	91.44%	.681
3.	.232	77	11.49%	1686	82.85%	.714
4.	.508	407	26.77%	2457	93.67%	.669
5.	.368	83	13.696%	2223	86.36%	.727
7.	.193	86	16.444%	1204	58.08%	.416
9.	.280	72	12.698%	1863	76.16%	.635
10.	.604	644	67.932%	2471	95.40%	.275
11.	.199	16	8.247%	1625	77.78%	.695
13.	.179	68	8.262%	1185	47.19%	.389
15.	.149	24	16.783%	1257	62.44%	.457
16.	.194	64	20.126%	1351	62.66%	.425
18.	.188	37	14.122%	1400	81.49%	.674
22.	.187	35	15.351%	1337	75.11%	.598

Table 4.1.14. (Continued)

Item no.	<i>p</i> -value	Num. Lower	Per. Lower	Num. Upper	Per. Upper	D-value
24.	.101	13	11.404%	852	80.75%	.694
26.	.079	17	11.333%	624	63.41%	.521
29.	.084	16	15.842%	638	70.03%	.542
32.	.177	91	24.011%	1207	67.13%	.431
34.	.101	3	4.762%	893	75.67%	.709

Note. 1. Num. Lower = Numbers of lower group, who answered item correctly.
 2. Per. Lower = Percentage of lower group, who answered item correctly.
 3. Num. Upper = Numbers of upper group, who answered item correctly.
 4. Per. Upper = Percentage of upper group, who answered item correctly.

According to Table 4.1.14, items 1, 2, 5, and 10 were identified as moderately difficult items, whereas items 3, 7, 9, 11, 13, 15, 16, 18, 22, 24, 26, 29, 32, 34 were identified as very difficult items. On the other hand, all items in Table 4.1.14 showed well discrimination (based on D-value), except item 10.

4.2. Mathematics Subtest (MS)

4.2.1 Descriptive Analysis

Table 4.2.1 represents the frequency distribution for the Mathematics Subtest. The sample of students for the MS was approximately evenly distributed with 5087(50.9%) male and 4913 (49.1) female students. There was no missing data for gender identification.

Table 4.2.1. *Frequency Distribution of Gender of Student for Mathematics Subtest*

Gender of Student	Number	Percent
Male	5087	50.9
Female	4913	49.1
Total	10000	100.0

The test mean score and the standard deviation were 9.17 and 8.40, respectively.

Skewness and kurtosis results show that the distribution was positively skewed and leptokurtic

(Skewness = 1.44, Kurtosis= 1.68). The standard error of measurement was 2.13. Cronbach`s alpha of the MS also was .94 for the total group.

Table 4.2.2 presents the item difficulty (p), the standard deviation of items, and item discriminations (r). The difficulty indices range from .600 to .054. The mean difficulty of the test was .229, which indicates that MS is highly difficult for examinees. Also, the mean discrimination of the test is .535, which shows the MS is moderately discriminating for examinees.

Table 4.2.2. *Descriptive Statistics for Mathematics Subtest Items*

Item No.	Item difficulty (p)	SD	Item Discrimination(r)
1.	.516	.500	.539
2.	.563	.496	.507
3.	.506	.500	.602
4.	.561	.496	.329
5.	.303	.460	.622
6.	.600	.490	.519
7.	.223	.416	.619
8.	.184	.387	.656
9.	.214	.410	.623
10.	.247	.431	.578
11.	.339	.473	.660
12.	.252	.434	.699
13.	.106	.308	.542
14.	.401	.490	.629
15.	.142	.349	.446
16.	.234	.423	.371
17.	.071	.257	.345
18.	.269	.444	.693
19.	.174	.379	.515
20.	.205	.404	.481
21.	.120	.325	.564
22.	.377	.485	.495
23.	.238	.426	.581
24.	.134	.341	.542
25.	.247	.431	.671
26.	.093	.291	.488
27.	.177	.382	.585
28.	.119	.324	.484
29.	.111	.314	.552

Table 4.2.2. (Continued)

Item No.	Item difficulty (p)	SD	Item Discrimination(r)
30.	.095	.293	.347
31.	.054	.226	.392
32.	.240	.427	.651
33.	.122	.327	.618
34.	.145	.352	.553
35.	.284	.451	.581
36.	.082	.274	.469
37.	.068	.251	.382
38.	.089	.284	.445
39.	.131	.337	.482
40.	.136	.343	.545

$N = 10,000$.

4.2.2. Cochran Mantel Haenszel Procedure (C-M-H)

Research Question 8: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?

The first research question in the study is related to the Cochran-Mantel-Haenszel method, which was conducted with the SAS 9.4 statistical software program. The C-M-H is used for detecting uniform DIF for the Mathematics Subtest items.

To implement the C-M-H method, the PROC FREQ procedure in SAS 9.4 was used, like the FMS analysis. If C-M-H p -value is less than a significant level ($p < .001$), and Breslow-Day Test for Homogeneity of the Odds Ratios' p -value is higher or equal than a significant level ($p \geq .001$), the item is indicating uniform DIF. *Odds Ratio* section in the C-M-H output helps to identify which item shows DIF for which gender. If the significant odds ratio is greater than one, the item shows DIF in favor of females, whereas if the odds ratio is less than one the item reveals DIF in favor of males (focal group=females, reference group= males) (focal group= 2, reference group= 1). Table 4.2.3 presents the results of the C-M-H procedure for the MS items.

Table 4.2.3. Results of Cochran-Mantel-Haenszel Analysis for Mathematics Subtest Items

Item no.	C-M-H <i>p</i> -value	C-M-H Odds ratio	Log Odds ratio	MH-DIF	Breslow- Day Test <i>p</i> -value	Breslow -Day Test χ^2	95% CI	Conclusion
1.	<.0001*	1.5951	0.4669	-1.09	0.0012**	27.37	1.44, 1.75	Uni. DIF
2.	<.0001*	1.3172	0.2755	-0.64	0.6030**	7.32	1.19, 1.45	Uni. DIF
3.	0.0004*	1.2150	0.1947	-0.45	0.8975**	4.20	1.09, 1.35	Uni. DIF
4.	0.4628	0.9683	-0.0322	-	0.1387	13.56	0.88, 1.05	No DIF
5.	<.0001*	1.5143	0.4149	-0.97	0.1745**	12.74	1.35, 1.69	Uni. DIF
6.	<.0001*	1.5653	0.4480	-1.05	0.1986**	12.26	1.40, 1.73	Uni. DIF
7.	<.0001*	1.3037	0.2652	-0.62	0.2479**	11.42	1.15, 1.47	Uni. DIF
8.	0.0349	0.8635	-0.1467	-	0.3097	10.52	0.75, 0.98	No DIF
9.	0.0040	1.1955	0.1785	-	0.00001	38.68	1.05, 1.35	No DIF
10.	0.8338	1.0120	0.0119	-	0.5277	8.06	0.90, 1.13	No DIF
11.	<.0001*	1.6541	0.5032	-1.18	0.2557**	11.29	1.47, 1.85	Uni. DIF
12.	0.0007*	1.2513	0.2241	-0.52	0.0030**	29.94	1.09, 1.42	Uni DIF
13.	0.9007	1.0097	0.0096	-	0.2759	10.99	0.86, 1.17	No DIF
14.	<.0001*	1.6607	0.5072	-1.19	0.1615**	13.02	1.48, 1.85	Uni. DIF
15.	<.0001*	0.6793	-0.3866	0.90	0.0817**	15.35	0.59, 0.77	Uni. DIF
16.	<.0001*	0.7564	-0.2791	0.65	0.0329**	18.19	0.68, 0.83	Uni. DIF
17.	0.4887	0.9444	-0.0572	-	0.0238	19.16	0.80, 1.11	No DIF
18.	<.0001*	1.3436	0.2953	-0.69	0.0889**	15.07	1.18, 1.52	Uni. DIF
19.	0.1687	0.9188	-0.0846	-	0.0890	15.07	0.81, 1.03	No DIF
20.	0.8053	0.9862	-0.0139	-	0.4331	9.04	0.88, 1.10	No DIF
21.	0.0461	0.8602	-0.1505	-	0.0676	12.98	0.76, 0.99	No DIF
22.	<.0001*	0.4693	-0.7565	1.77	0.5343**	7.99	0.42, 0.51	Uni. DIF
23.	0.7984	1.0148	0.0146	-	0.7305	6.09	0.90, 1.13	No DIF
24.	0.0895	1.1258	0.1184	-	0.0097	21.74	0.98, 1.29	No DIF
25.	0.0019	1.2157	0.1953	-	0.2791	10.95	1.07, 1.37	No DIF
26.	0.1245	0.8855	-0.1216	-	0.3258	10.31	0.75, 1.03	No DIF
27.	0.0526	0.8828	-0.1246	-	0.0080	22.28	0.77, 1.00	No DIF
28.	0.8240	0.9846	-0.0155	-	0.2344	11.63	0.85, 1.12	No DIF
29.	0.6468	0.9656	-0.0350	-	0.1985	12.27	0.83, 1.12	No DIF
30.	<.0001*	0.5173	-0.6591	1.54	0.6256**	7.11	0.44, 0.59	Uni. DIF
31.	<.0001*	0.5946	-0.5198	1.22	0.1958**	12.32	0.49, 0.72	Uni. DIF
32.	<.0001*	0.6555	-0.4223	0.99	0.3968**	9.45	0.57, 0.74	Uni. DIF

Table 4.2.3. (Continued)

Item no.	C-M-H <i>p</i> -value	C-M-H Odds ratio	Log Odds ratio	MH-DIF	Breslow-Day Test <i>p</i> -value	Breslow-Day Test χ^2	95% CI	Conclusion
33.	<.0001*	0.7018	-0.3541	0.83	0.2884**	10.81	0.60, 0.81	Uni. DIF
34.	<.0001*	0.7635	-0.2698	0.63	0.0035**	24.55	0.66, 0.87	Uni. DIF
35.	<.0001*	0.8017	-0.2210	0.51	0.6968**	6.42	0.71, 0.89	Uni. DIF
36.	0.0175	0.8211	-0.1971	-	0.0113	19.74	0.69, 0.96	No DIF
37.	<.0001*	0.7055	-0.3488	0.81	0.0164**	20.26	0.59, 0.83	Uni. DIF
38.	0.0200	0.8344	-0.1810	-	0.0384	17.73	0.71, 0.97	No DIF
39.	<.0001*	0.5907	-0.5264	1.23	0.0780**	15.50	0.51, 0.67	Uni. DIF
40.	<.0001*	0.7349	-0.3082	0.72	0.8431**	4.89	0.64, 0.84	Uni. DIF

Note. 1. *p*-value =.001. Uni. DIF =uniform DIF.
 2. DF for C-M-H is 1 and DF for the Breslow-Day test is 9.
 3. If the C-M-H *p*-value is <.001*, and the Breslow-Day test *p*-value is \geq .001**, the item reveals uniform DIF.

Research Question 8 Response: Based on the C-M-H results, items 1, 2, 3, 5, 6, 7, 11, 12, 14, 15, 16, 18, 22, 30, 31, 32, 33, 34, 35, 37, 39, and 40 show evidence of uniform DIF. Therefore, 55% of the 40 items are identified as exhibiting uniform DIF. Items 1, 2, 3, 5, 6, 7, 11, 12, 14, and 18 favor female examinees, whereas items 15, 16, 22, 30, 31, 32, 33, 34, 35, 37, 39, and 40 favor males.

Table 4.2.4 presents the items by DIF in the ETS Delta Scale based on the DIF levels.

Table 4.2.4. *The Items with DIF Categorization in the ETS Delta Scale*

	Item numbers favoring female examinees	Item numbers favoring male examinees
Category A (negligible)	2, 3, 5, 7, 12, 18	15, 16, 33, 34, 35, 37, 40
Category B (moderate)	1, 6, 11, 14	31, 32, 39
Category C (large)	-	22, 30

4.2.3. Logistic Regression Procedure

Research Question 9: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is defined as having uniform and non-uniform gender DIF using the Logistic Regression method?

The second research question is associated with the logistic regression method. The purpose of using the logistic regression method in this study was to identify uniform and non-uniform DIF and to compare the results with the C-M-H method results for the Mathematics Subtest items. The results are evaluated based on the Likelihood Ratio Test. To interpret logistic regression results, first, the *p-value* for the interaction should be examined for evidence of non-uniform DIF. If it is not statistically significant ($p > .001$), the *p-value* for the main effect in the model with gender and total score should be checked for evidence of uniform DIF. If *p-values* for both interaction and main effect are not statistically significant ($p > .001$), finally, there is no DIF in the item. After identifying items with non-uniform and uniform DIF, like the C-M-H procedure, the *Odds Ratio* table helps to clarify which item reveals DIF for which gender. Table 4.2.5 presents the results of the Logistic regression procedure for MS items.

Table 4.2.5. Results of Logistic Regression Analysis for the Mathematics Subtest Items

Item no.	Model 1 χ^2	Model 2 χ^2	Model 3 χ^2	<i>p</i> -value for main effect	<i>p</i> -value for interaction	Odds Ratio for Gender	Log Odds Ratio	MH- DIF	Conclusion
1.	4068.942	4160.437	4171.347	0.00000*	0.00096***	1.61	0.47	-1.11	Non-uni. DIF
2.	3714.343	3744.142	3746.025	0.00000*	0.16997**	1.30	0.26	-0.61	Uni. DIF
3.	5626.364	5639.841	5645.222	0.00008*	0.02035**	1.22	0.19	-0.46	Uni. DIF
4.	1243.530	1243.914	1247.235	0.15684	0.06838	0.97	-0.03	-	No DIF
5.	4224.696	4278.987	4281.294	0.00000*	0.12878**	1.51	0.41	-0.96	Uni. DIF
6.	4332.222	4401.576	4390.936	0.00000*	0.00111**	1.53	0.42	-0.99	Uni. DIF
7.	3706.479	3726.157	3733.244	0.00000*	0.00777**	1.31	0.27	-0.63	Uni. DIF
8.	3996.408	4001.033	3997.626	0.54389	0.0649	0.85	-0.16	-	No DIF

Table 4.2.5. (Continued)

Item no.	Model 1 χ^2	Model 2 χ^2	Model 3 χ^2	p -value for main effect	p -value for interaction	Odds Ratio for Gender	Log Odds Ratio	MH-DIF	Conclusion
9.	3709.080	3717.935	3733.126	0.00001*	0.0001***	1.20	0.18	-0.42	Non-uni. DIF
10.	3261.029	3261.082	3261.243	0.89842	0.68801	1.01	0.009	-	No DIF
11.	5274.642	5348.133	5354.330	0.00000*	0.0128**	1.65	0.50	-1.17	Uni. DIF
12.	5294.618	5305.195	5318.313	0.00001	0.00029***	1.24	0.21	-0.50	Non-uni. DIF
13.	2301.177	2301.457	2303.283	0.34877	0.17649	1.04	0.03	-	No DIF
14.	5168.069	5252.006	5255.682	0.00000*	0.05518**	1.66	0.50	-1.19	Uni. DIF
15.	1637.610	1672.094	1645.747	0.01710	0.00000	0.68	-0.38	-	No DIF
16.	1251.827	1280.296	1284.049	0.00000*	0.05273**	0.76	-0.27	0.64	Uni. DIF
17.	892.053	892.147	893.1908	0.56641	0.3070	0.97	-0.03	-	No DIF
18.	5328.158	5348.747	5348.847	0.00003*	0.75221**	1.34	0.29	-0.68	Uni. DIF
19.	2294.073	2295.519	2298.972	0.08636	0.06313	0.92	-0.08	-	No DIF
20.	2070.400	2070.421	2070.633	0.88986	0.64476	0.99	-0.01	-	No DIF
21.	2569.283	2572.311	2569.306	0.98866	0.08302	0.87	-0.13	-	No DIF
22.	2672.330	2909.502	2848.517	0.00000*	0.0000***	0.47	-0.75	1.77	Non-uni. DIF
23.	3266.412	3266.485	3266.614	0.90397	0.72021	1.01	0.009	-	No DIF
24.	2409.038	2413.028	2417.466	0.01478	0.03515	1.15	0.13	-	No DIF
25.	4704.837	4714.248	4707.610	0.25003	0.00998	0.19	-1.66	-	No DIF
26.	1815.189	1816.656	1815.251	0.96933	0.23582	0.90	-0.10	-	No DIF
27.	3036.930	3040.296	3037.343	0.81348	0.08573	0.88	-0.12	-	No DIF
28.	1864.455	1864.462	1864.455	0.9998	0.93573	1.00	0	-	No DIF
29.	2413.845	2413.869	2414.659	0.66553	0.37398	0.98	-0.02	-	No DIF
30.	938.574	1015.903	992.997	0.00000*	0.0000***	0.51	-0.67	1.58	Non-uni. DIF
31.	1092.626	1117.316	1101.482	0.01194	0.00007	0.60	-0.51	-	No DIF
32.	4298.801	4347.75	4331.270	0.00000*	0.00005***	0.64	-0.44	1.04	Non-uni. DIF
33.	3140.725	3159.521	3149.572	0.01199	0.00161	0.70	-0.35	-	No DIF
34.	2561.822	2576.759	2571.065	0.00984	0.01703	0.76	-0.27	-	No DIF
35.	3480.389	3497.543	3495.922	0.00042*	0.2030**	0.79	-0.23	0.55	Uni. DIF
36.	1642.102	1646.044	1642.630	0.76809	0.06464	0.84	-0.17	-	No DIF
37.	1072.928	1086.548	1076.530	0.16509	0.00155	0.72	-0.32	-	No DIF
38.	1501.286	1505.430	1504.471	0.20344	0.32747	0.84	-0.17	-	No DIF
39.	1883.308	1942.122	1925.230	0.00000*	0.00004***	0.58	-0.54	1.28	Non-uni. DIF
40.	2446.796	2465.210	2462.555	0.00038*	0.10324**	0.73	-0.31	0.73	Uni. DIF

Note. 1. DF for p -value for the main effect is 2 and DF for p -value for interaction is 1.

2. if p -value for main effect is $\geq .001$, the item reveals No DIF.

3. If p -value for main effect is $<.001^*$, and p -value for interaction is $>.001^{**}$, the item shows Uniform DIF.
4. If p -value for main effect is $<.001^*$, and p -value for interaction is $<.001^{***}$, the item reveals Non-Uniform DIF.

Research Question 9 Response: Based on the logistic regression results, items 2, 3, 5, 6, 7, 11, 14, 16, 18, 35, and 40 indicate uniform-DIF. Items 1, 9, 12, 22, 30, 32, and 39, indicate non-uniform-DIF. Therefore, 45% of the 40 items are identified as DIF. Items 1, 2, 3, 5, 6, 7, 9, 11, 12, 14, and 18 favor female examinees, whereas items 16, 22, 30, 32, 35, 39, and 40 favor males. Table 4.2.6 presents the items by DIF in the ETS Delta Scale.

Table 4.2.6. The Items with DIF Categorization in the ETS Delta Scale

	Item numbers favoring female examinees	Item numbers favoring male examinees
Category A (negligible)	2, 3, 5, 6, 7, 9, 12, 18	16, 35, 40
Category B (moderate)	1, 11, 14	32, 39
Category C (large)	-	22, 30

Research Question 10: Do the Cochran-Mantel-Haenszel and Logistic Regression technique results match each other in identifying gender DIF for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Research Question 10 Response: The C-M-H is not sensitive for detecting non-uniform DIF. Therefore, when comparing the two methods results, which items indicate DIF can be considered. Based on table 4.2.7, both methods detect DIF in the same items, except items 9, 15, 31, 33, 34, and 37.

Table 4.2.7. Comparison of Types of DIF based on Two Chi-square Methods

Methods	Items with Uniform DIF	Items with Non-Uniform DIF
<i>C-M-H</i>	1, 2, 3, 5, 6, 7, 11,12, 14, 15, 16, 18, 22, 30, 31, 32, 33, 34, 35, 37, 39, and 40.	–
<i>Logistic Regression</i>	2, 3, 5, 6, 7, 11, 14, 16, 18, 35, and 40.	1, 9, 12, 22, 30, 32, and 39.

Note. The bold items are the favor of females.

4.2.4 2-PL IRT-LR Procedure

4.2.4.1. Checking Model Assumptions and Clarifying Which Model is Better for The Test

Research Question 11: Are the IRT assumptions met for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination data?

Research Question 11 Response: To evaluate the dimension of latent factors, *eigenvalues of the Polychoric Correlation Matrix* tables for each gender group were provided by the PROC IRT procedure. The tables show that there is only one dominant eigenvalue identified with 20.514 (the second eigenvalue is 3.05) for males and 20.663 (the second eigenvalue is 2.82) for females in the model, which supports model unidimensional.

According to the SAS/STAT 14.3® User Guide Book, independency of observed responses (items) is proof of the local independence assumption (p. 4828). Besides, Table 4.2.8 presents model fit statistics based on the models in the IRT-LR.

Table 4.2.8. Model Fit Statistics for MS

	Rasch Model	1-PL	2-PL	3-PL	4-PL
<i>Log Likelihood</i>	- 150914.4027	-150914.4033	-147900.4602	-146196.7762	-146012.3133
<i>AIC</i>	301992.80543	301992. 80669	296120.92045	292873.55234	292664.62657
<i>BIC</i>	302584.05334	302584.0546	297274.57491	294604.03403	294971.93549

Note.1. $p < .001$.

2. AIC= Akaike`s information criterion (smaller is better).

3. BIC= Bayesian information criterion (smaller is better).

To make a decision about which model is better fit, Log-Likelihood (LL), AIC, and BIC criteria were considered. When comparing 2-, 3- and 4- PL IRT-LR models, the smaller log-likelihood value was for the 2-PL IRT-LR model. So, the 2-PL IRT-LR procedure was used for the MS, like the Fundamental Mathematics Subtest. During the following IRT-LR methods in the study, log-likelihood values for each parameter were compared to detect differential item functioning.

Research Question 12: How do the difficulty, and discrimination parameter estimations compare between male and female students for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Research Question 12 Response: PROC IRT provides the *Item Parameter Estimates* table, including difficulty and slopes estimates, standard errors, and p-values for each item. The difficulty parameter refers to item difficulty (b parameter), and the slope parameter refers to item discrimination (a parameter) in the IRT procedure. In table 4.2.9, the range of difficulty and discrimination parameter estimates were presented based on the groups. The discrimination ranges for both groups include positive values in this study, so all the items (responses) are adequate measures of the latent trait. Also, the discrimination results support that the test is more discriminating for females than males.

For the male examinees, most of the difficulty parameters are higher than 0, which suggests that most of the items in this test are relatively hard. Besides, for female students, the difficulty parameters have higher estimates than male’s difficulty parameter estimates.

Table 4.2.9. *Item Parameter Estimate Ranges for Each Group*

Group	Discrimination (a) parameters range	Difficulty (b) parameters range
Male	0.47 to 1.62	-0.41 to 3.14
Female	0.38 to 1.64	-0.45 to 3.65

In table 4.2.10, item discrimination (a) and item difficulty (b) parameter estimates are presented separately for both male and female examinees.

Table 4.2.10. *Item Parameter Estimate for Each Group*

Item no.	b parameter for male	a parameter for male	b parameter for female	a parameter for female
1.	-0.008	1.009	-0.21	1.07
2.	-0.20	1.008	-0.31	0.99
3.	-0.06	1.52	-0.09	1.52
4.	-0.41	0.47	-0.37	0.39
5.	0.68	1.14	0.54	1.14
6.	-0.27	1.22	-0.45	1.13
7.	1.02	1.04	0.93	1.08
8.	1.04	1.21	1.18	1.20
9.	1.07	0.99	0.97	1.14
10.	0.91	0.95	0.99	0.91
11.	0.48	1.48	0.34	1.49
12.	0.72	1.46	0.67	1.64
13.	1.86	0.85	1.90	0.88
14.	0.30	1.39	0.14	1.39
15.	1.89	0.59	2.11	0.66
16.	1.44	0.49	2.24	0.38
17.	3.14	0.51	2.96	0.59
18.	0.64	1.62	0.61	1.52
19.	1.36	0.83	1.72	0.67
20.	1.36	0.71	1.45	0.69
21.	1.65	0.90	1.75	0.96
22.	0.18	0.77	0.80	0.75
23.	0.93	0.99	1.02	0.95
24.	1.70	0.83	1.65	0.88
25.	0.73	1.50	0.79	1.24
26.	2.08	0.77	2.23	0.77
27.	1.23	0.95	1.36	0.97
28.	1.86	0.78	2.03	0.73
29.	1.75	0.90	1.82	0.93
30.	2.49	0.52	3.65	0.44
31.	2.82	0.63	3.08	0.70
32.	0.72	1.20	0.97	1.21
33.	1.42	1.12	1.65	1.15
34.	1.43	0.90	1.75	0.83

Table 4.2.10. (Continued)

Item no.	b parameter for male	a parameter for male	b parameter for female	a parameter for female
35.	0.65	1.01	0.89	0.92
36.	2.24	0.74	2.38	0.78
37.	2.80	0.58	3.01	0.62
38.	2.20	0.72	2.52	0.67
39.	1.63	0.76	2.23	0.68
40.	1.50	0.90	0.81	0.81

Research Question 13: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?

Two-Parameter Logistic Analysis via IRT-LR Using SAS 9.4 program.

The research question 13 is associated with the two-parameter logistic model using the IRT-Likelihood Ratio test, and its ability to detect differences between groups while considering the examinee's ability, item discrimination and item difficulty parameters.

To implement the 2-PL IRT-LR method, the PROC IRT procedure in SAS 9.4 was used. The results are interpreted based on the Log-Likelihood (LL) values, which are general model fit LL, freely estimated intercepts` LL, freely estimated intercept and slopes` LL (*constrained baseline method*). For each type of DIF, *p-values* were computed.

Table 4.2.11 presents the results of the 2-PL IRT-LR analysis of the MS items. To conduct the analyses, *p-value* for ab-DIF should be looked first to determine statistical significance ($p < .001$). If the *p-value* for ab-DIF is less than significant level ($p < .001$), the item including non-uniform DIF. If *p-value* for ab-DIF is not less than at significance level ($p < .001$) and if the *p-value* for b-DIF is less than significant level ($p < .001$), the item shows evidence of

uniform DIF. If *p-value* for b-DIF is less than 0, and *p-value* for ab-DIF is $>.001$, the item reveals No DIF.

Table 4.2.11. *Results of 2-PL IRT-LR Analysis for Mathematics Subtest Item*

Item no.	Intercept LL	Intercept and Slope LL	p-value for b-DIF	p-value for ab-DIF	Conclusion
1.	-147919.81	-147919.83	0.0002*	0.870**	Uniform DIF
2.	-147905.00	-147905.40	0.175	0.526	No DIF
3.	-147901.09	-147901.11	0.882	0.865	No DIF
4.	-147901.85	-147904.36	0.271	0.112	No DIF
5.	-147907.87	-147908.45	0.046	0.444	No DIF
6.	-147909.45	-147914.06	0.003	0.031	No DIF
7.	-147901.31	-147902.22	0.623	0.340	No DIF
8.	-147904.67	-147906.65	0.102	0.159	No DIF
9.	-147900.94	-147904.70	0.236	0.052	No DIF
10.	-147900.93	-147901.66	0.750	0.390	No DIF
11.	-147909.41	-147910.27	0.020	0.353	No DIF
12.	-147900.84	-147903.44	0.393	0.106	No DIF
13.	-147901.72	-147901.80	0.719	0.783	No DIF
14.	-147913.56	-147913.67	0.004	0.739	No DIF
15.	-147925.89	-147926.42	0.00001*	0.465**	Uniform DIF
16.	-147913.40	-147922.82	0.00005*	0.002**	Uniform DIF
17.	-147903.18	-147903.32	0.413	0.712	No DIF
18.	-147902.40	-147902.56	0.550	0.687	No DIF
19.	-147900.55	-147910.46	0.018	0.001	No DIF
20.	-147901.18	-147901.50	0.791	0.572	No DIF
21.	-147907.25	-147907.76	0.062	0.477	No DIF
22.	-148004.59	-148006.45	0.000*	0.172**	Uniform DIF
23.	-147900.82	-147902.09	0.652	0.260	No DIF
24.	-147900.71	-147901.01	0.907	0.585	No DIF
25.	-147903.59	-147909.07	0.034	0.019	No DIF
26.	-147903.40	-147904.50	0.256	0.294	No DIF
27.	-147905.31	-147906.13	0.128	0.365	No DIF
28.	-147900.63	-147902.14	0.640	0.219	No DIF
29.	-147902.68	-147903.03	0.462	0.556	No DIF

Table 4.2.11. (Continued)

Item no.	Intercept LL	Intercept and Slope LL	p-value for b-DIF	p-value for ab-DIF	Conclusion
30.	-147929.45	-147945.38	0.00000***	0.00007***	Non-Uniform DIF
31.	-147918.75	-147922.62	0.00006**	0.049**	Uniform DIF
32.	-147923.27	-147926.93	0.00001**	0.055**	Uniform DIF
33.	-147911.80	-147917.28	0.00077**	0.019**	Uniform DIF
34.	-147906.15	-147912.95	0.005	0.009	No DIF
35.	-147908.40	-147913.65	0.004	0.021	No DIF
36.	-147906.54	-147907.65	0.066	0.291	No DIF
37.	-147911.98	-147913.76	0.004	0.182	No DIF
38.	-147902.82	-147906.35	0.116	0.059	No DIF
39.	-147920.51	-147935.69	0.00000***	0.0001***	Non-Uniform DIF
40.	-147906.63	-147915.41	0.001	0.003	No DIF

Note. 1. $p < .001$.

2. General Log likelihood = -147900.4602.

2. if p -value for b-DIF is < 0 , and p -value for *ab-DIF* is $> .001$, the item reveals No DIF.

3. If p value for b-DIF is $< .001^*$, and p -value for *ab-DIF* is $> .001^{**}$, the item reveals Uniform DIF.

4. If p -value for b-DIF is $< .001^*$, and p -value for *ab-DIF* is $< .001^{***}$, the item reveals Non-Uniform DIF.

Research Question 13 Response: Based on 2-PL IRT-LR results, items 1, 15, 16, 22, 31, 32, and 33 indicate uniform DIF. Also, items 30 and 39 are flagged as non-uniform DIF. As a result, it can be said that 22.5% of the 40 items are identified DIF.

Nine items were flagged for DIF in the MS subtest. To check which items favor which gender, parameter b can be compared because parameter b refers to item difficulty (Odett, 1997). Table 4.2.12 presents the comparison of significant differences between manifest groups on the MS items using the 2-PL IRT-LR model.

Table 4.2.12. Comparison of Significant Differences between Manifest Groups on MS Items Using 2-PL IRT-LR Model

Test Items by DIF	Females Parameter "b"	Males Parameter "b"	Difference in the "b" parameter
1.	-0.21	-0.008	0.202
15.	2.11	1.89	-0.22
16.	2.24	1.44	-0.8
22.	0.8	0.18	-0.62
30.	3.65	2.49	-1.16
31.	3.08	2.82	-0.26
32.	0.97	0.72	-0.25
33.	1.65	1.42	-0.23
39.	2.23	1.63	-0.6

$p < .001$.

According to Table 4.2.12, all items with DIF favored males because their b parameter differences are negative, except item 1, which favored females. However, to understand which item with non-uniform DIF favored which gender, it needs to check the items in the ability scale. Therefore, items with non-uniform DIF, which are 30 and 39 were evaluated based on the ICC (see Appendix A, figure A.2.). According to ICCs, items 30 and 39 favored high ability group, which is reference (male) group.

Research Question 14: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination showed DIF using all three methods?

Research Question 14 Response: The final research question in the study is associated with comparing non-IRT, and IRT approaches result based on how many items reveal DIF in their results. In table 4.2.13, a comparison of the three methods is presented.

Table 4.2.13. Comparison of Types of DIF based on Non-IRT and IRT-LR Methods

Methods	Items with Uniform DIF	Items with Non-Uniform DIF	Percentage of DIF
<i>C-M-H</i>	1, 2, 3, 5, 6, 7, 11, 12, 14, 15, 16, 18, 30, 22, 31, 32, 33, 35, 34, 37, 39, and 40.	–	55 %
<i>Logistic Regression</i>	2, 3, 5, 6, 7, 11, 14, 16, 18, 35, and 40.	1, 9, 12, 22, 30, 32, and 39.	45 %
<i>2-PL IRT-LR</i>	1,15, 16, 22, 31, 32, and 33.	30 and 39	22.5%

$p < .001$.

After DIF analysis, the items with DIF need to be compared using their p -value and D -values to draw conclusions. Table 4.2.14 presents the conclusion of the items, which are including DIF or not, based on the two-group approach.

Table 4.2.14. The Conclusion of the Items, which are including DIF or not, based on the Two Groups Approach.

Item no.	p -value	Num. Lower	Per. Lower	Num. Upper	Per. Upper	D-Value
1.	.516	522	18.69%	2521	92.78%	.741
2.	.563	495	29.46%	2492	94.86%	.654
3.	.506	206	16.81%	2585	96.49%	.797
5.	.303	124	9.39%	2003	80.70%	.713
6.	.600	473	35.97%	2588	96.24%	.602
7.	.223	90	12.95%	1611	82.44%	.695
9.	.214	86	16.04%	1552	83.48%	.674
11.	.339	88	9.91%	2222	87.06%	.772
12.	.252	48	8.43%	1944	86.59%	.782
14.	.401	156	29.71%	2380	93.92%	.642
15.	.142	86	11.04%	889	54.37%	.433
16.	.234	294	23.57%	1150	54.19%	.306
18.	.269	40	11.29%	2005	84.77%	.735
22.	.377	223	17.64%	1922	81.64%	.640
29.	.111	16	7.11%	845	66.01%	.589
30.	.095	64	7.53%	572	29.12%	.216
31.	.054	19	8.48%	367	38.38%	.299
32.	.240	67	12.64%	1772	83.94%	.713
33.	.122	12	4.09%	1031	55.82%	.517

Table 4.2.14. (Continued)

Item no.	<i>p</i> -value	Num. Lower	Per. Lower	Num. Upper	Per. Upper	D-Value
34.	.145	41	18.38%	1043	72.03%	.536
35.	.284	74	25.96%	1808	82.67%	.567
37.	.068	23	4.80%	452	33.28%	.285
38.	.089	13	9.77%	632	61.84%	.521
39.	.131	44	12.29%	926	48.15%	.359

Note. 1. Num. Lower = Numbers of lower group, who answered item correctly.
 2. Per. Lower = Percentage of lower group, who answered item correctly.
 3. Num. Upper = Numbers of upper group, who answered item correctly.
 4. Per. Upper = Percentage of upper group, who answered item correctly.

According to Table 4.2.14, items 1, 2, 3, 6, 14, 22 were identified as moderately difficult items, whereas items 5, 7, 9, 11, 12, 15, 16, 18, 29, 30, 31, 32, 33, 34, 35, 37, 38, and 39 were identified as very difficult items. On the other hand, all items in Table 4.2.14 showed good discrimination (based on D-value), except items 16, 30, 31 and 37.

Chapter 5

Discussion

This chapter presents the summary, findings, conclusions, and the implications of the study.

5.1. Summary

Psychometric properties of tests cover reliability, validity, and fairness. As nationwide examinations, it is expected that the two tests examined in the present study should have high reliability, validity, and fairness. Differential item functioning analyses were used to evaluate the validity of the two nationwide exams. There are multiple methods that can be employed to detect differential item functioning. In classical test theory, student performances are evaluated based on test scores. Therefore, the results of CTT approaches are test-dependent. However, for item response theory, student performances are assessed based on student abilities; that is why IRT approaches give test-independent results. There are multiple ways to investigate DIF in classical test and item response theories. The purpose of this study was to use Cochran-Mantel-Haenszel and Logistic Regression as CTT approaches, and 2-PL IRT-LR was used as an IRT approach, to evaluate gender DIF for two nationwide exams in Turkey.

Before conducting DIF analysis, descriptive statistics were analyzed for both subtests. According to the results, the Fundamental Mathematics subtest (FMS) is very difficult (mean item difficulty is .356) and moderately discriminating (mean item discrimination is .554) for

students. Similarly, the Mathematics subtest (MS) is very difficult (mean item difficulty is .229) and moderately discriminating (mean item discrimination is .535).

To investigate items with DIF, non-IRT approaches were conducted first, and then the IRT approach was conducted for both subtests. To classify test items based on topics, table 5.1.1 was used.

Table 5.1.1. *General Mathematics Subtopics*

Number	Arithmetic	Algebra	Geometry	Advanced Math
Four operations	Percentage	Functions	Plane geometry	Permutation
Integers	Ratio-Proportion	Equations	Co-ordination	Combination
Digits	Profit-Loss	Graphs	Trigonometry	Probability
Sets and Subsets	Average	Polynomial		

5.2. Findings and Conclusions for Cochran-Mantel- Haenszel Analysis

In this part, Research question 1 and 8 are discussed together because both research questions are related to Cochran-Mantel-Haenszel analysis.

Research Question 1: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?

Research Question 8: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform gender DIF using the Cochran-Mantel-Haenszel method?

5.2.1 Findings for Fundamental Mathematics and Mathematics Subtests based on C-M-H

These research questions were examined using the Cochran-Mantel-Haenszel method to detect the differences between male and female examinees in the Fundamental Mathematics and Mathematics subtests. Each test has 40 items, and the gender differences were tested at the .001 significant level (p -value).

For the Fundamental Mathematics subtest, 18 out of 40 items (45%) were identified as DIF. Half of the items (50%) favored male examinees, and the other half (50%) favored female examinees. When looking at the FMS items that favored females, the items divided into three mathematics subtopics, which are *number* (items 2, 4, 5, 9, 11, 13), *algebra* (items 15 and 16), and *geometry* (item 34). On the other hand, the FMS items that favored males, also divided into three mathematics subtopics, which are *arithmetic* (item 3, 7, 10, 18, 22, and 24), *advanced math* (item 26 and 29), and *geometry* (item 32).

Based on the ETS delta scale, item 5, 11, and 18 were in category C, which means large DIF. On the other hand, items 4, 9, 10, 15, 24, 26, and 29 were in category B, which is moderate, and the other items were in category A, which means negligible DIF.

For the Mathematics subtest, according to the C-M-H results, 22 out of 40 items (55%) revealed DIF. About 45.4 % of the items favored female examinees, 54.6 % of the items favored male examinees. When looking at the MS items, which favored females, the items divided into three mathematics subtopics, which are *number* (items 1, 2, 3, 5, and 6), *arithmetic* (item 12), and *algebra* (items 7, 11, 14, 18). On the other hand, the MS items that favored males also divided into two mathematics subtopics, which were *advanced math* (items 15, 16, and 22), and *geometry* (items 30, 31, 32, 33, 34, 35, 37, 39, and 40). Using the ETS delta scale, items 22 and

30 were in category C, whereas, items 1, 6, 11, 14, 31, 32, and 39 were in category B, and the other items were in category A.

5.3. Findings and Conclusions for Logistic Regression Analysis

In the second part of the DIF analysis, research question 2 and research question 9 are discussed due to their link with Logistic Regression analysis.

Research Question 2: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is characterized as having uniform and non-uniform gender DIF using the Logistic Regression method?

Research Question 9: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is defined as having uniform and non-uniform gender DIF using the Logistic Regression method?

5.3.1. Findings for Fundamental Mathematics and Mathematics Subtests based on LR

The second research question was answered using the Logistic Regression method to detect DIF for male and female examinees in the Fundamental Mathematics and Mathematics subtests. For the Fundamental Mathematics subtest, 16 out of 40 items (40%) were identified as DIF. All items with DIF in the Logistic regression procedure agreed with items with DIF based on the C-M-H procedure. There were only two items in the C-M-H procedure (items 13 and 32) that disagreed with the results of the Logistic regression procedure. However, according to the C-M-H results, both test items were in Category A based on the ETS delta scale. Therefore, these disagreements between the two non-IRT approaches are minor. In the FMS, there were more items with non-uniform DIF than the items with uniform DIF: 9 (56.2 %) and 7 (43.8%), respectively. The FMS items that favored females divided into three mathematics subtopics,

which are *number* (items 2, 4, 5, 9, 11), *algebra* (items 15 and 16), and *geometry* (item 34). In contrast, the FMS items that favored males divided into two mathematics subtopics, which are *arithmetic* (items 3, 7, 10, 18, 22, and 24) and *advanced math* (items 26 and 29). Based on the logistic regression method, there were no geometry items favoring males in the FMS test.

Based on the ETS delta scale categorization, the logistic regression results agree with Cochran-Mantel-Haenszel results in Category C, Category B, and Category A, except item 29. According to the LR results, item 29, which favored males was in Category C instead of Category B.

For the Mathematics Subtest, 22 out of the 40 items (55%) are identified with DIF. Compare to the FMS, MS had more disagreements between the C-M-H and LR methods. Although the Logistic regression method indicated that item 9 revealed DIF, C-M-H method did not identify these items as items with DIF. In contrast, C-M-H method indicated that items 15, 31, 33, 34 and 37 revealed DIF, but logistic regression method did not identify these items with DIF. Due to the higher sensitivity to detect uniform and non-uniform DIF, logistic regression results are more acceptable than the C-M-H results.

Furthermore, 11 out of the 18 (61%) items favored females, whereas 7 (38.9 %) out of 18 items favored male students. In the MS, there were more items with uniform DIF than non-uniform DIF, which are 11 (61.1 %) and 7 (38.9 %), respectively.

The MS items that favored females divided into three mathematics subtopics, which are *number* (items 1, 3, 5 and 6), *arithmetic* (item 12) and *algebra* (items 7, 9, 11, 14, 18). On the other hand, the MS items that favored males divided into two mathematics subtopics, *advanced math* (items 16, 22, and 35), and *geometry* (items 30, 32, 39, and 40). In the ETS delta scale,

both methods agreed that item 22 and item 30 are in Category C, and there is no item, which favored females, with significant DIF.

5.4. Conclusion for Fundamental Mathematics and Mathematics Subtests for Non-IRT Analysis

To compare C-M-H results with LR results, research question 3 and research question 10 were asked for both subtests.

Research Question 3: Do the Cochran-Mantel-Haenszel and Logistic Regression technique results for DIF match each other in the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Research Question 10: Do the Cochran-Mantel-Haenszel and Logistic Regression technique results match each other in identifying gender DIF for the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Firstly, the Cochran-Mantel-Haenszel procedure was used to investigate differential item functioning for gender. The obtained significant results were considered to get conclusions about bias at the item-level.

For the FMS, items 2, 4, 5, 9, 11, 13, and 15, which favored females, are a type of number questions, except item 15 (*algebra-polynomial*) and they require four operations. In contrast, items 3, 7, 10, 18, 22, 24, which favored males, are *arithmetic* questions, and they require problem-solving abilities. Although previous studies show that males are better than females for geometry, one geometry item favored both males and one geometry item favored females, item 34 and item 32, respectively, in the FMS test. However, both items are in Category A regarding the ETS delta scale, which means acceptable DIF. In addition, items 26 and 29

avored males that were types of advanced math. These results support that males tend to outperform females in application and analysis levels on the FMS.

Interestingly, even though the items with DIF in the MS had a similar conclusion with the FMS test items with DIF, the geometry items 30, 31, 32, 33, 34, 35, 37, 39, and 40 favored males. However, there is only one item (30) is in Category C, which means having large DIF.

Another remarkable result for the FMS involved items 5 and 11, which favored females, had large DIF (category C) and both items included geometric shapes, and their topics were related to *number*. Item 18 in the FMS had large DIF, which favored males and was related to *arithmetic*.

In light of the ETS delta scale results, the MS subtest has no items in category C for the favored females. Also, for the favored male examinees, there are only two items that are in category C, which is item 22 and item 30. These items are related to *advanced math* (item 22) and *geometry* (item 30).

In the second part of the analysis, the logistic regression procedure was used to investigate DIF for gender. For the FMS, the LR results are consistent with the C-M-H results (88.8% agreement), whereas for the MS subtest, the LR results were compatible with the C-M-H results at an 86.3 % level of agreement.

5.5. Findings and Conclusions for 2-PL IRT-LR Analysis

In the final part of the DIF analysis discussion, research question 4 and research question 11 are discussed first to check assumptions for both subtests, and then difficulty and discrimination parameter estimations are conducted with research questions 5 and 12 to find differences between female and male examinees.

Research Question 5: How do the difficulty, and discrimination parameter estimations compare between male and female students in the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination?

Research Question 12: How do difficulty, and discrimination parameter estimations compare between male and female students in the Mathematics subtest of the MSPC - 2018 Higher Education Institutions Examination?

In the FMS, the most items were relatively hard for male and female examinees because the IRT item difficulty parameters are higher than 0. There were only items 2, 4, and 10 identified as easy items for males, whereas items 4 and 10 were identified as easy items for females. On the other hand, discrimination ranges for both groups suggest that all the items (responses) are adequate measures of latent traits.

In the MS, most items were relatively hard for male and female examinees because the IRT item difficulty parameters are higher than 0. There were only items 1,2, 3, 4, and 6 identified as easy items for both males and females.

Research Question 6: What percentage of the items on the Fundamental Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?

Research Question 13: What percentage of the items on the Mathematics Subtest of the MSPC - 2018 Higher Education Institutions Examination is identified as having uniform and non-uniform gender DIF using the 2-PL IRT-LR method?

Item response theory uses a different approach compared to classical test theory to detect DIF, but this approach shares the same matching criterion variable, which is *ability* instead of *the*

total test score. Therefore, IRT eliminates some challenges of CTT approaches, like the use of observed variables, and gives more reliable results than CTT.

5.5.1. Findings for 2-PL IRT-LR analysis

Under research questions 6 and 13, to conduct 2-PL IRT-LR test, a constrained baseline method was used for both subtests. For DIF analysis, log-likelihood ratio values were used.

For the Fundamental Mathematics Subtest, 10 out of the 40 items (25%) were flagged with DIF. The items with DIF were divided into three mathematics subtopics: *arithmetic* (items 1, 3, 7, 10, 18, 22, and 24), *advanced math* (items 26 and 29), and *geometry* (item 32). The items in the FMS, which are 1, 7, 22, and 32 exhibited uniform DIF, whereas, items 3, 10, 18, 24, 26, and 29 revealed non-uniform DIF.

Gender DIF in these items were identified by examining differences in the “b” parameters (i.e., item difficulty). If the difference is negative, the item favored males. If the difference is positive, the item favored females. For the 10 items with DIF, no item favored females because all difference values were negative. On the other hand, for items with non-uniform DIF were evaluated based on the ability scales. According to ICCs, all items with non-uniform DIF favored males, except item 10.

For the Mathematics Subtest, 9 out of the 40 items (22.5%) were flagged with DIF. For the nine items with DIF, only item 1 favored females because the difference value was negative, whereas the other eight items favored male students.

Item 1, which favored females, is related to the *number* mathematics subtopic. The other items, which favored males can be divided into two mathematics subtopics, which are *advanced math* (items 15, 16, and 22), and *geometry* (item 30, 31, 32, 33, and 39). Except for item 30 and

item 39, the items were flagged as showing non-uniform DIF. Therefore, according to ICCs, items 30 and 39 favored males.

5.5.2 Findings based on Two-Group Approach

For FMS, after analyzing the items with DIF, item 10 was moderately difficult and not a well discriminating item based on the two-group approach.

For MS, after analyzing the items with DIF, items 16, 30, 31, and 37 were very difficult and not a well discriminating items based on two-group approach.

Therefore, item 10 in FMS and item 16, 30, 31, and 37 were categorized as items, which require revisiting.

5.5.3. Conclusions for 2-PL IRT-LR Analysis and Discussion between non-IRT and IRT Approaches

Broad Research Question 1.1. For each test, what percentage of the items show gender DIF?

For the Fundamental Mathematics subtest, 18 (45%), 16 (40%), and 10 (25%) out of 40 items were identified as DIF in C-M-H, LR, and 2-PL IRT-LR analysis, respectively. For the Mathematics subtest, 22 (55%), 18 (45%), and 9 (22.5%) out of 40 items were identified as DIF in C-M-H, LR, and 2-PL IRT-LR analysis, respectively.

Broad Research Question 1.2. To what extent is there agreement in the identification of gender DIF using these 3 methods, which are Cochran-Mantel-Haenszel, Logistic Regression, and 2-PL IRT-LR?

According to the 2-PL IRT-LR analysis for both subtests, the results are similar to the non-IRT approaches in terms of subtopics of items, which favored females and males. Table 5.5.1. presents all methods comparisons based on subtopics of items, which favored male or females.

Table 5.5.1. All methods` Comparisons based on Subtopics of Items, which favor males or females.

Methods	FMS		MS	
	Males	Females	Males	Females
Cochran-Mantel-Haenszel	Arithmetic (3,7,10,18, 22, 24) Advanced Math (26, 29) Geometry (32)	Number (2, 4, 5, 9, 11,13) Algebra (15,16) Geometry (34)	Advanced Math (15,16, 22) Geometry (30, 31, 32, 33, 34, 35, 37, 39, 40)	Number (1, 2, 3, 5, 6) Arithmetic (12) Algebra (7,11,14,18)
Logistic Regression	Arithmetic (3,7,10,18, 22, 24) Advanced Math (26, 29)	Number (2, 4, 5, 9,11) Algebra (15,16) Geometry (34)	Advanced Math (16, 22, 35) Geometry (30,32, 39, 40)	Number (1, 2, 3, 5, 6) Arithmetic (12) Algebra (7, 9, 11,14,18)
2-PL IRT-LR	Arithmetic (1,3,7,18, 22, 24) Advanced Math (26, 29) Geometry (32)	Arithmetic (10)	Advanced Math (15,16, 22) Geometry (30, 31,32, 33, 39)	Number (1)

Note. 1. FMS= Fundamental Mathematics Subtests.
 2. MS= Mathematics Subtests.
 3. Based on ETS delta scale, bold, italic, and underlined item numbers in parenthesis refers to effect sizes of DIF that is in Category A (negligible), Category B (moderate), and Category C (large), respectively.

In previous studies, males tended to outperform females in visual (Abedalaziz, 2010) and spatial skills (Baran-Cohen, 2005; Geary, 1996; Halpern et al., 2007). However, according to table 5.5.1, there was no solid evidence to substantiate a conclusion that males are better than females in terms of visual and spatial skills in the FMS. One geometry item (item 32) favored males based on the C-M-H and 2 PL IRT-LR results. According to the ETS delta scale, which is used for the C-M-H results, item 32 was in Category A, which means negligible DIF. Moreover,

item difficulty parameters differences, which is used for an item with uniform DIF in the IRT result, were not significant (-0.22).

For MS, although all DIF methods identified several geometry items that favored male students, there is only item 30 in Category C (large DIF) based on non-IRT approaches and had a significant item parameter difference (-1.16) based on the IRT approach. However, after the two-groups approach, item 30 was identified as very difficult and not discriminating. Therefore, the same conclusion is reached with the FMS.

On the other hand, there was some solid evidence to substantiate a conclusion that females tended to outperform in four operation skills and numerical abilities (Abedalaziz, 2010; Cepni, 2011). Because for both tests, items with *number* subtopic favored females. Also, the items with *advanced math and algebra* subtopics favored males. It supports that males are better than females in problem-solving skills and analytical thinking abilities (Cepni, 2011). In addition, although *arithmetic* items in both tests favored both male and female examinees, it can be said that these items favored males because item 10 in the FMS needs to be revisited (not discriminating well item) and item 12 in the MS was in Category A (negligible DIF).

Broad Research Question 1.3. To what extent is there agreement the identification of uniform and non-uniform DIF using these 3 methods?

The Logistic Regression method and 2- PL IRT-LR method can be compared based on DIF types, which are uniform DIF and non-uniform DIF. C-M-H is not designed to detect non-uniform DIF.

For the FMS subtest, there were no agreement between the items with uniform DIF. However, items 18, 24, and 29 are flagged as non-uniform DIF in both methods. The main

difference between two methods occurs in item 26 because while the LR method reveals item 26 as uniform, 2-PL IRT-LR shows as non-uniform DIF.

For the MS subtest, there was only agreement on item 16 with uniform DIF, and items 30 and 39 with non-uniform DIF. The main differences between two methods occurred in items 1, 22 and 32 because while the LR method reveals these items as non-uniform, 2-PL IRT-LR shows as uniform DIF.

5.6. Recommendations for Future Research

Based on these findings, the following recommendations can be considered for future studies:

1. Conduct further research including additional variables besides gender, especially age, and region.
2. Conduct and compare 2-PL IRT-LR and 3-PL IRT-LR for this kind of large case data.
3. Compare the methods in terms of Type 1 error rate and power.
4. Cognitive Interviewing may be recommended after DIF analysis to evaluate items with DIF.

References

- Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal*, 5, 101-116.
- American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing Performances (Type I Error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory and Practice*, 14(6), 2186-2193.
- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.
- Baron-Cohen, S. (2005, January). The essential difference: The male and female brain. In *Phi Kappa Phi Forum* (Vol. 85, No. 1, pp. 23-26). National Forum: Phi Kappa Phi Journal.
- Berberoğlu, G. (1995). Differential item functioning analysis of computation, word problem, and geometry questions across gender and SES groups. *Studies in Educational Evaluation*, 21, 439-456.
- Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

- Cees, A. & Glas, W. Retrieved from van der Linden, W. J. (Eds.). (2018). *Handbook of modern item response theory. Volume two*. Springer Science & Business Media.
- Çepni, Z. (2011). Değişen madde fonksiyonlarının sibtest, Mantel Haenzsel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi. *Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yayınlanmamış Doktora Tezi*.
- Chen, Y., Cao, C., & Green, S. (2014). Field Test Analysis Report: SAS Macro and Item/Distractor/DIF Analyses. Retrieved from <https://www.slideserve.com/naiara/field-test-analysis-report-sas-macro-and-item-distractor-dif-analyses>
- Choi, J. (2017). A Review of PROC IRT in SAS. *Journal of Educational and Behavioral Statistics*, 42(2), 195-205.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and practice*, 17(1), 31-44.
- DeMars, C. (2010). *Item response theory*. Oxford University Press.
- De Ayala, R. J. (2008). *The theory and practice of item response theory*. Guilford Publications.
- Gamerman et. al. Retrieved from van der Linden, W. J. (Eds.). (2018). *Handbook of modern item response theory. Volume three*. Springer Science & Business Media.
- Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, 19(2), 229-247.
- Geary, D. C. (1999). Sex differences in mathematical abilities: Commentary on the math-fact retrieval hypothesis.
- Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, 8(1), 1-51.

- Hambleton, R. K., & Swaminathan, H. (1985). 1985: *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.
- Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on a Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates, Inc.
- Kalaycıoğlu, D. B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim*, 36(161).
- Kamata, A., & Vaughn, B. K. (2004). An Introduction to Differential Item Functioning Analysis. *Learning Disabilities: A Contemporary Journal*, 2(2), 49-69.
- Lopez, G. E. (2012). Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and logistic regression procedures.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748.
- Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, 16(2), rm2.
- Matlock Cole, K., & Peak, I. (2017). PROC IRT: A SAS procedure for item response theory. *Applied Psychological Measurement*, 41(4), 311-320.
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2), 127-143.

- Odett, D. C. (1997). An evaluation of item response theory for detecting differential item functioning of examinees' responses to the seventh-grade mathematics MEAP test investigating learner characteristics. UMI Number: 9815355.
- ÖSYM. (August 2018). 2018 YKS Değerlendirme Raporu / 2018 HEIE Evaluation Report. Report no:9.
- ÖSYM. (2018). 2018 YKS Kilavuz / 2018 HEIE Guide Book. Retrieved from:
https://dokuman.osym.gov.tr/pdfdokuman/2018/YKS/KILAVUZ_28062018.pdf
- ÖSYM. Website: <https://www.osym.gov.tr/TR,15134/2018-yks-tyt-ayt-ydt-temel-soru-kitapciklari-ve-cevap-anahtarlari.html>.
- Özer, M. (2018). Ölçme, Seçme ve Yerleştirme Merkezinin Stratejik Hedefleri ve Yeni Yönelimleri. *Journal of Higher Education & Science/Yükseköğretim ve Bilim Dergisi*, 8(2).
- Penny, T. Using the SAS" System to Detect Differential Item Functioning. *Statistics, Data Analyzing, and Modeling*. University Research Associates, Jamestown, NC.
- Philip, A., & Ojo, B. O. (2017). Application of item characteristic curve (ICC) in the selection of test items. *British Journal of Education*, 5(2), 21-41.
- Popham, W. J. (1999). *Classroom assessment: What teachers need to know*. Allyn & Bacon, A Viacom Company, 160 Gould St., Needham Heights, MA 02194; World Wide Web: <http://www.abacon.com>.
- Royse, D., Thyer, B. A., & Padgett, D. K. (2009). *Program evaluation: An introduction*. Cengage Learning.
- SAS Institute Inc. 2017. SAS/STAT® 14.3 User's Guide. *Proc IRT Procedure*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2013. SAS/STAT® 13.1 User's Guide. *The FREQ Procedure*. Cary, NC: SAS Institute Inc.

Selkow, P. (1985). Male/female differences in mathematical ability: A function of biological sex or perceived gender role? *Psychological Reports*, 57(2), 551-557.

Sireci , S. G. & Rios, J. A., (2013) Decisions that make a difference in detecting differential item functioning, *Educational Research and Evaluation*, 19:2-3, 170-187, DOI: 10.1080/13803611.2013.767621

Şenferah, S. (2015). 2010 Seviye belirleme sınavı matematik alt testi için değişen madde fonksiyonlarının ve madde yanlılığının incelenmesi. *Yayınlanmamış Yüksek Lisans Tezi. Gazi Üniversitesi. Eğitim Bilimleri Enstitüsü. Ankara.*

Wang, W.-C. & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods.

Yıldırım, H. (February 2015). An investigation of item bias of mathematics subtest in the 2012-year Level Determination Exam. Master`s thesis. Gazi University. Educational Science Institutes. Department of Measurement and Evaluation in Education.

Yurdagül, H. & Aşkar, P. (2004). Ortaöğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı'nın cinsiyete göre madde yanlılığı açısından incelenmesi. *Eğitim Bilimleri ve Uygulama Dergisi*, 3(5), 3-20

Zhang, Y. (2015). Multiple ways to detect differential item functioning in SAS. In *Proceedings of SAS Global Forum 2015 Conference* (pp. 2900-2015).

Appendices

Appendix A: Item Characteristic Curves

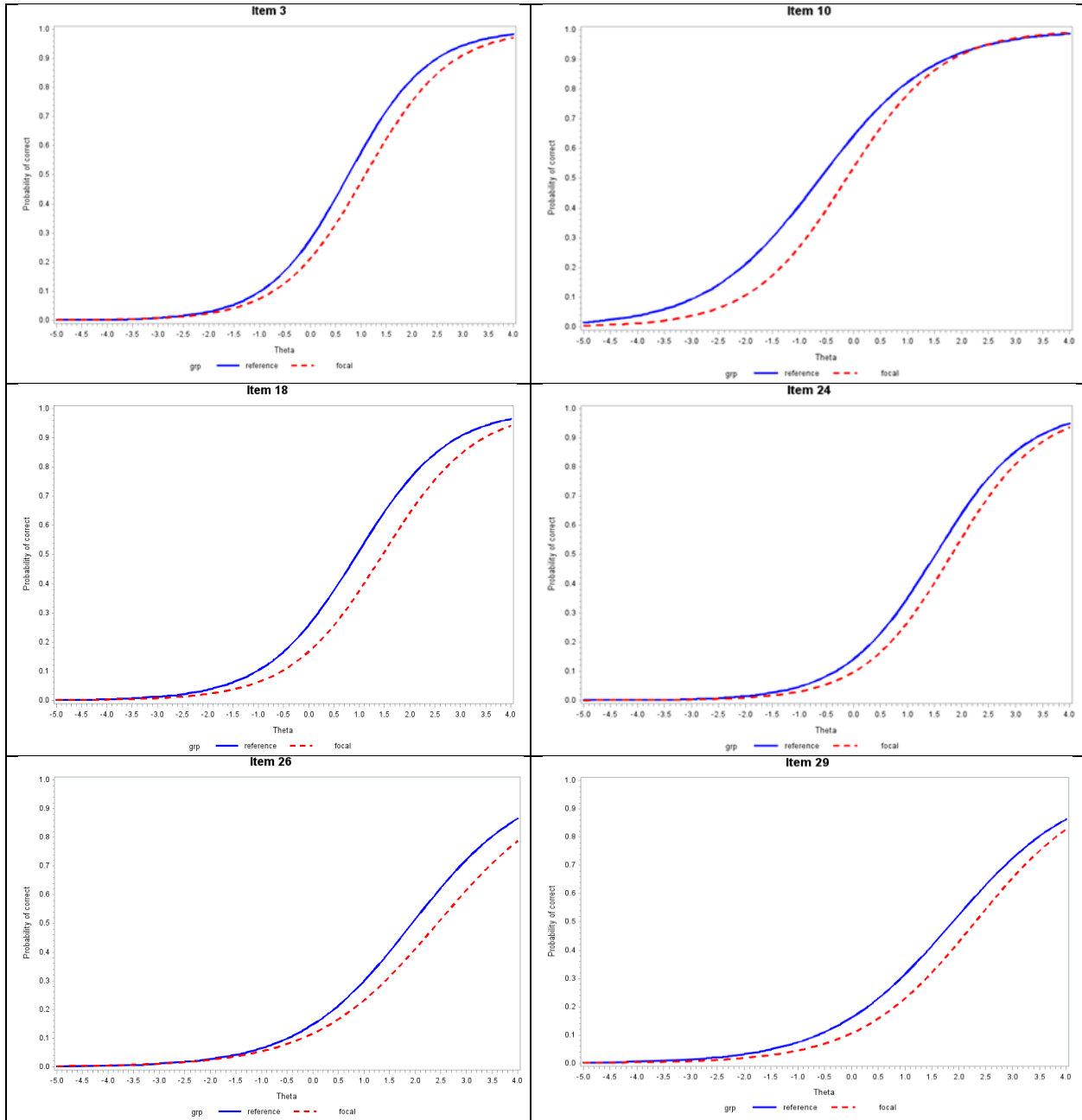


Figure A.1. Item Characteristic Curves for Items with Non-Uniform DIF in the FMS.

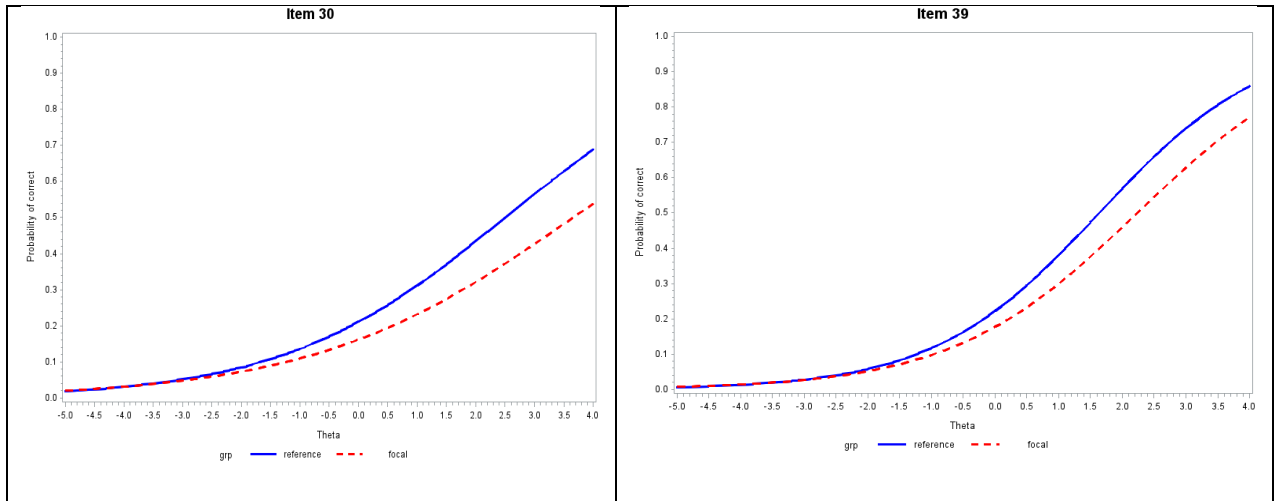


Figure A.2. Item Characteristic Curves for Items with Non-Uniform DIF in the MS.

Appendix B: Some Original and Translated Test Items in the MSPC- 2018 HEIE

IMPORTANT NOTICE

According to the Law on Intellectual and Artistic Works (*Turkish name: Fikir ve Sanat Eserleri Kanunu*) in Turkey, "All rights of these test items used in this thesis belong to the MSPC (ÖSYM) in Turkey. For whatever purpose, copying, photographing, reproduction of all or reproduction of any part of it in any way cannot be done without the written permission of the MSPC (ÖSYM)."

Table B.1., the FMS items, which are identified with DIF in all methods, presented in original (Turkish) and translated (English) languages. Original items were taken from ÖSYM website, whereas the items were translated by a private translation office in Turkey.

Table B.1. *The FMS Items, which is Identified with DIF in All Methods.*

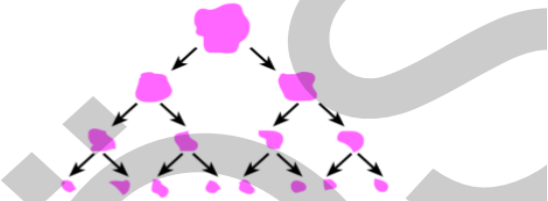
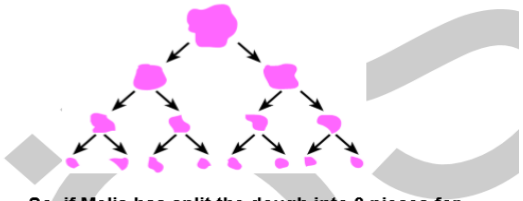
Original Item	Translated Item
<p>2. Eline bir oyun hamuru alan Melis, şekilde gösterildiği gibi her adımda elindeki her bir oyun hamurunu 2 parçaya ayırıyor ve 3. adım sonunda 8 parça oyun hamuru elde ediyor.</p>  <p>Melis başlangıçtan itibaren her adımda, elindeki her bir oyun hamurunu 2 yerine 3 parçaya ayırsaydı 4. adım sonunda kaç parça oyun hamuru elde ederdi? A) 12 B) 36 C) 51 D) 72 E) 81</p>	<p>2. Melis has a piece of play dough. She splits the dough into two pieces for each step as below. On the 3rd step, she has 8 pieces.</p>  <p>So, if Melis has split the dough into 3 pieces for each step instead of 2, how many pieces would she have at the end of the 4th step? A) 12 B) 36 C) 51 D) 72 E) 81</p>

Table B.1. (Continued)


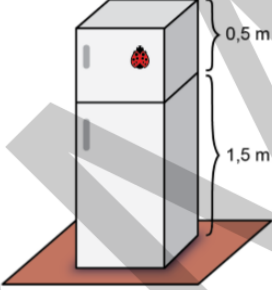
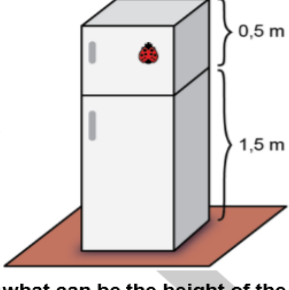
Original Item	Translated Item
<p>3. İki bölmeli dikdörtgenler prizması şeklindeki bir buzdolabının alt bölümü 1,5 metre, üst bölümü ise 0,5 metre yüksekliğindedir. Buzdolabının üst bölümünün üzerine  şeklindeki bir süs aşağıdaki gibi yapıştırılıyor.</p>	<p>3. There is a fridge with two compartments in the shape of a rectangular prism. The height of the bottom compartment is 1.5 m, the top compartment is 0.5 m. An ornament in the shape of a lady bug is put on the fridge as below.</p>
	
<p>Buna göre, yapıştırılan bu süsün yerden yüksekliği metre türünden aşağıdakilerden hangisi olabilir? A) $\sqrt{2}$ B) $\sqrt{3}$ C) $\sqrt{5}$ D) $\sqrt{6}$ E) $\sqrt{7}$</p>	<p>So, what can be the height of the ornament from the floor? A) $\sqrt{2}$ B) $\sqrt{3}$ C) $\sqrt{5}$ D) $\sqrt{6}$ E) $\sqrt{7}$</p>
<p>4. I. $\begin{bmatrix} -2 & \square & 2 \end{bmatrix}$ II. $\begin{bmatrix} 2 & \square & -2 \end{bmatrix}$ III. $\begin{bmatrix} -2 & \square & -2 \end{bmatrix}$ İfadelerindeki boş kutuların içine toplama (+), çıkarma (-) ve çarpma (\times) sembolleri hangi sırayla yerleştirilirse üç işlemin sonucu da aynı sayıya eşit olur? A) $\begin{matrix} \text{I} & \text{II} & \text{III} \\ + & \times & - \end{matrix}$ B) $\begin{matrix} - & + & \times \end{matrix}$ C) $\begin{matrix} - & \times & + \end{matrix}$ D) $\begin{matrix} \times & + & - \end{matrix}$ E) $\begin{matrix} \times & - & + \end{matrix}$</p>	<p>4. I. $\begin{bmatrix} -2 & \square & 2 \end{bmatrix}$ II. $\begin{bmatrix} 2 & \square & -2 \end{bmatrix}$ III. $\begin{bmatrix} -2 & \square & -2 \end{bmatrix}$ Which of the symbols should be put in the empty boxes so that they all give out the same result? A) $\begin{matrix} \text{I} & \text{II} & \text{III} \\ + & \times & - \end{matrix}$ B) $\begin{matrix} - & + & \times \end{matrix}$ C) $\begin{matrix} - & \times & + \end{matrix}$ D) $\begin{matrix} \times & + & - \end{matrix}$ E) $\begin{matrix} \times & - & + \end{matrix}$</p>

Table B.1. (Continued)

















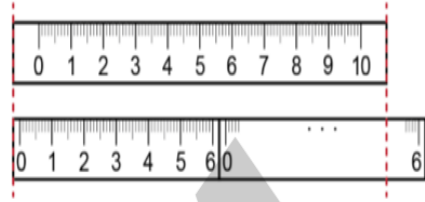
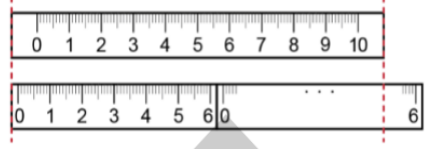
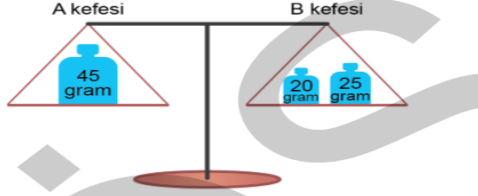
Original Item	Translated Item
<p>5. n kenarlı bir düzgün çokgenin içine yazılan bir a doğal sayısıyla oluşturulan sembol ile $n \cdot a^n$ sayısı gösterilmektedir.</p> <p>Örneğin,  sembolü ile $3 \cdot 2^3 = 24$ sayısı gösterilmektedir.</p> <p>Buna göre,</p> <p> \cdot </p> <p>çarpımının değerini gösteren sembol aşağıdakilerden hangisidir?</p> <p>A)  B)  C) </p> <p>D)  E) </p>	<p>5. The number of $n \cdot a^n$ is symbolized in a regular polygon with n number of edges.</p> <p>For example, the symbol of  represents $3 \cdot 2^3 = 24$</p> <p>So, what does the symbol below</p> <p> \cdot </p> <p>equal to in terms of symbols?</p> <p>A)  B)  C) </p> <p>D)  E) </p>
<p>7. Her iki tarafında da 0,8 cm mesafe olan 10 cm'lik bir cetvelin altına, her iki tarafında da 0,2 cm mesafe olan 6 cm'lik özdeş iki cetvel, aralarında boşluk bırakılmadan uç uca birleştirilerek şekildeki gibi soldan hizalanmıştır.</p>  <p>Buna göre, 10 cm'lik cetvelin sağ kenarı 6 cm'lik cetvelin hangi noktasıyla hizalanmıştır?</p> <p>A) 4 B) 4,5 C) 4,8 D) 5 E) 5,2</p>	<p>7. One ruler of 10 cm has 0.8 cm distance on both of its sides. Two rulers of 6 cm which have 0.2 cm distance on both of their sides are put together end to end as below.</p>  <p>So, what does the right end of the ruler of 10 cm correspond to on the ruler of 6 cm?</p> <p>A) 4 B) 4,5 C) 4,8 D) 5 E) 5,2</p>
<p>9. a, b ve c pozitif tam sayıları için $a(b + c)$ ifadesi bir tek sayıya eşittir.</p> <p>Buna göre,</p> <p>I. $a^b + c$ II. $b^c + a$ III. $c^a + b$</p> <p>ifadelerinden hangileri her zaman tek sayıya eşittir?</p> <p>A) Yalnız II B) Yalnız III C) I ve II D) II ve III E) I, II ve III</p>	<p>9. a, b, and c are positive integers. $a(b + c)$ The figure above equals to an odd number.</p> <p>So,</p> <p>I. $a^b + c$ II. $b^c + a$ III. $c^a + b$</p> <p>11. which of the figures above is <u>always</u> an odd number?</p> <p>A) Only II B) Only III C) I and II D) II and III E) I, II, and III</p>

Table B.1. (Continued)

Original Item

Translated Item

10. Üzerlerinde kütleleri yazılı olan ağırlıklar, eşit kollu bir terazinin kefelerine şekildeki gibi yerleştirilerek terazi dengelenmiştir.

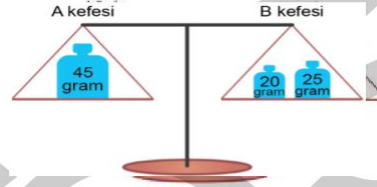


Aşağıda verilen ağırlıklardan biri terazinin B kefesine eklenip B kefesindeki ağırlıklardan biri A kefesine aktarıldığında bu terazi yine dengede kalmaktadır.



- Buna göre, bu işlem sırasında B kefesine eklenen ağırlık kaç gramdır?
A) 10 B) 15 C) 30 D) 35 E) 40

10. The weights below have their masses. They have been placed on two sides of a scale as below.

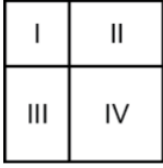


When one of the weights below is placed on the B Scale and one of the weights from B Scale is transferred to A Scale, the scale is still balanced.



- So, which one of the weights is placed on B Scale?
A) 10 B) 15 C) 30 D) 35 E) 40

11. Kenar uzunluğu a birim olan bir kare, şekildeki gibi dört bölgeye ayrıldığında I numaralı bölge kenar uzunluğu b birim olan bir kare belirtmektedir.



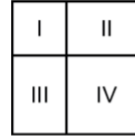
Bu koşulu sağlayan her a ve b sayısı için

$$a^2 - 2ab + 2b^2$$

ifadesi hangi iki bölgenin alanları toplamına eşittir?

- A) I ve II B) I ve IV C) II ve III
D) II ve IV E) III ve IV

11. The square below has edges of a unit. The area labeled as I has edges of b unit when the square is split into four areas as below.



For every a and b that ensures that condition,

$$a^2 - 2ab + 2b^2$$

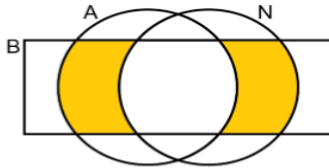
what does the figure above equal to?

- A) I and II B) I and IV C) II and III
D) II and IV E) III and IV

13. Aşağıdaki Venn şemasında

- A harfi ile başlayan isimler kümesi A,
- N harfi ile biten isimler kümesi N,
- 5 harfli isimler kümesi B

ile gösterilmiştir.



Buna göre,

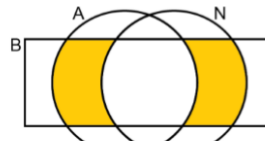
$K = \{\text{AÇELYA, AHMET, AYSUN, BEREN, KENAN, NERMİN}\}$

kümesinin elemanlarından kaç tanesi şekildeki boyalı bölgeler ile gösterilen kümenin elemanıdır?

- A) 1 B) 2 C) 3 D) 4 E) 5

13. The Venn diagram below represents the below statements.

- The names starting with A are shown in the set A
- The names ending with N are shown in the set N
- The names with 5 letters are shown in set B.



So,

$K = \{\text{AÇELYA, AHMET, AYSUN, BEREN, KENAN, NERMİN}\}$

how many of the elements of the set above are members of the areas of yellow color?

- A) 1 B) 2 C) 3 D) 4 E) 5

Table B.1. (Continued)

Original Item

Translated Item

15. $P(x)$ bir polinom olmak üzere, $P(a) = 0$ eşitliğini sağlayan a sayısına bu polinomun bir kökü denir.

$P(x)$ ve $R(x)$ polinomları için

$$P(x) = x^2 - 1$$

$$R(x) = P(P(x))$$

eşitlikleri veriliyor.

Buna göre,

- I. -1
- II. 0
- III. 1

sayılarından hangileri $R(x)$ polinomunun köküdür?

- A) Yalnız I B) Yalnız II C) Yalnız III
D) I ve III E) II ve III

15. $P(x)$ is a polynomial. The a number a that provides the equation of $P(a) = 0$ is called a root of this polynomial. For the polynomials of $P(x)$ and $R(x)$, the following equations are given.

$$P(x) = x^2 - 1$$

$$R(x) = P(P(x))$$

So,

- I. -1
- II. 0
- III. 1

Which of those are the roots of the $R(x)$ polynomial?

- A) Only I B) Only II C) Only III
D) I and III E) II and III

16. Türkiye'deki 81 ilin tamamını kapsayan bir projede; önce her bir ile p tane park yapılması, sonra da yapılan her bir parka a tane ağaç dikilmesi planlanmıştır.

Fakat, bu planda yapılacak park ve dikilecek ağaç sayısı yeterli bulunmamış ve önce her bir ile yapılması planlanan park sayısından 1 fazla sayıda park yapılmış, sonra da yapılan her bir parka dikilmesi planlanan sayıdan 1 fazla sayıda ağaç dikilmiştir.

Buna göre, son durumda dikilen toplam ağaç sayısı ile başlangıçta dikilmesi planlanan toplam ağaç sayısı arasındaki fark aşağıdakilerin hangisinde doğru olarak verilmiştir?

- A) 162
B) $81 \cdot a \cdot p$
C) $81 \cdot (a + p)$
D) $81 \cdot (a \cdot p + 1)$
E) $81 \cdot (a + p + 1)$

16. In a project that covers 81 cities of Turkey, it was planned to plant a amount of trees in p amount of parks.

However, the number of parks and trees were not enough as per this project. So, for each city, one more park has been established and for each park, one more tree has been planted.

So, which of the below shows the difference between total number of trees planted and number of trees that was planned to be planted?

- A) 162
B) $81 \cdot a \cdot p$
C) $81 \cdot (a + p)$
D) $81 \cdot (a \cdot p + 1)$
E) $81 \cdot (a + p + 1)$

18. Belirli bir bölgede ev ve arsa alım satım işlemi yapan Ali Bey'in bu işlemlerde kullandığı birim fiyatlar tabloda verilmiştir.

	Alış fiyatı (TL)	Satış fiyatı (TL)
Ev ($1 m^2$)	3000	3200
Arsa (1 dönüm)	20 000	25 000

Ali Bey, 450 000 TL'ye aldığı bir evin satışından elde ettiği paranın tamamı ile bir arsa almış ve sonra bu arsayı da satmıştır.

Buna göre, Ali Bey'in bu arsa satışından elde ettiği kâr kaç TL'dir?

- A) 90 000 B) 105 000 C) 110 000
D) 120 000 E) 125 000

18. Mr. Ali is working as a real-estate agent. The table below shows the prices that he buys and sells houses.

	Purchase Price (TL)	Selling Price (TL)
House ($1 m^2$)	3000	3200
Land (1 decaire)	20 000	25 000

Mr. Ali has purchased a house at 450 000 TL and bought a land with the whole of the profit he got from the sale of that house. Then, he sold the land.

So, what s Mr. Ali's profit from the land he sold?

- A) 90 000 B) 105 000 C) 110 000
D) 120 000 E) 125 000

Table B.1. (Continued)

Original Item	Translated Item
<p>22. Bir ayakkabı fabrikasında üretilen her bir ayakkabının A ve B standartlarına göre belirlenen numara değerleri arasında doğrusal bir ilişki bulunmaktadır.</p> <p>Bu fabrikada üretilen en küçük ayakkabının numara değeri A standardında 34, B standardında 7; en büyük ayakkabının numara değeri ise A standardında 46, B standardında 13'tür.</p> <p>Buna göre, B standardında numara değeri 11,5 olan bir ayakkabının, A standardındaki numara değeri kaçtır?</p> <p>A) 43 B) 42 C) 41 D) 40 E) 39</p>	<p>22. There is linear relationship between size values of every shoe in a shoe factory according to A and B standards.</p> <p>The smallest shoe is size 34 in A standard, 7 in B standard. The biggest shoe is size 46 in A standard, 13 in B standard.</p> <p>So, what is the size of a shoe in A standard which is size 11.5 in B standard?</p> <p>A) 43 B) 42 C) 41 D) 40 E) 39</p>
<p>24. Arif bir tarifte, yaş mısırın kurutulduğunda ağırlığının % 20 oranında azaldığını, kurutulmuş mısırın ise patlatıldığında ağırlığının % 10 oranında azaldığını okumuştur. Sonra, bu oranlara uygun olarak 720 gram patlamış mısır elde etmek için yeterli miktarda yaş mısır satın almıştır.</p> <p>Arif, aldığı yaş mısırın tamamını kurutup patlattıktan sonra istediği miktardan daha az patlamış mısır elde etmiş ve bu durumun tarifteki bir hatadan kaynaklandığını, % 20 olarak yazılan oranın aslında % 30 olması gerektiğini fark etmiştir.</p> <p>Buna göre, Arif'in elde ettiği patlamış mısır miktarı kaç gramdır?</p> <p>A) 630 B) 640 C) 660 D) 680 E) 690</p>	<p>24. In a recipe, Arif learned that fresh corn loses 20% of its weight when dried, and dry corn loses 10% of its weight when popped. Then, following these ratios, he bought enough fresh corn to make 720 grams of popcorn.</p> <p>After drying and popping the corn he bought, Arif got less popcorn than he desired. And he realized that this happened because of a mistake in the recipe. 20% should have been 30%.</p> <p>So, how many grams of popcorn did Arif get?</p> <p>A) 630 B) 640 C) 660 D) 680 E) 690</p>
<p>26. Bir açılışa katılan 25 davetlinin her biri için mandalina suyu, nar suyu ve portakal suyunun her birinden birer bardak hazırlanmış ve davetlilere ikram edilmiştir. İkram edilen bu içeceklerle ilgili aşağıdakiler bilinmektedir.</p> <ul style="list-style-type: none"> • Tüm davetliler en az bir çeşit içecek almıştır. • Aynı çeşit içecekten birden fazla bardak alan davetli <u>bulunmamaktadır</u>. • Yalnızca iki çeşit içecek alan davetli <u>bulunmamaktadır</u>. <p>Açılış sonunda 7 bardak mandalina suyu, 8 bardak nar suyu ve 9 bardak portakal suyunun <u>alınmadığı</u> belirlenmiştir.</p> <p>Buna göre, bu açılıшта üç çeşit içecek alan davetli sayısı kaçtır?</p> <p>A) 7 B) 9 C) 11 D) 13 E) 15</p>	<p>26. For each of 25 guest in an opening, one cup of mandarin juice, pomegranate juice, and orange juice, and those were served to the guests. Information about the juices is given below.</p> <ul style="list-style-type: none"> • Every guest received at least one juice. • <u>No guests</u> received more than one cup of the same type of juice. • <u>No guests</u> received only two types of juices. <p>After the end, 7 cups of mandarin juice, 8 cups of pomegranate juice, and 9 cups of orange juice have <u>left</u>.</p> <p>So, how many guests have received three types of juices in the opening?</p> <p>A) 7 B) 9 C) 11 D) 13 E) 15</p>

Table B.1. (Continued)

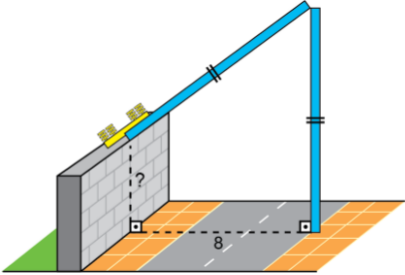
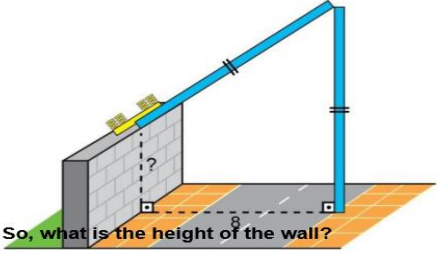
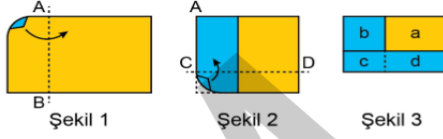
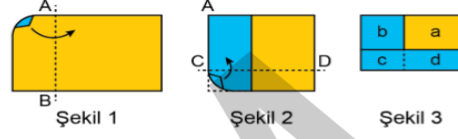
Original Item	Translated Item
<p>29. Bir elektronik tartı; her ölçümde, üzerine konulan ağırlığı % 20 olasılıkla gerçek ağırlığından 1 kilogram fazla, % 30 olasılıkla gerçek ağırlığından 1 kilogram az, % 50 olasılıkla da doğru tartmaktadır.</p> <p>Gerçek ağırlıkları sırasıyla 80 ve 81 kilogram olan Ali ile Mehmet bu tartıda birer kere tartılacaklardır.</p> <p>Buna göre, ölçüm sonunda Ali ile Mehmet'in ağırlıklarının birbirine eşit çıkma olasılığı yüzde kaçtır?</p> <p>A) 40 B) 35 C) 30 D) 25 E) 20</p>	<p>29. An electronic scale measures the weights on it as follows:</p> <p>1 kg more than real weight with 20% probability 1 kg less than real weight with 30% probability Real weight with 50% probability</p> <p>Ali and Mehmet, whose real weights are 80 and 81 respectively, will measure their own weights on this scale.</p> <p>What is the percentage probability of their weights being measured as equal?</p> <p>A) 40 B) 35 C) 30 D) 25 E) 20</p>
<p>32. Uzunluğu 20 metre olan mavi renkli elektrik direği, fırtına nedeniyle tam ortadan kırılmış ve direğin uç noktası şekilde görüldüğü gibi direğe 8 metre uzaklıkta bulunan duvarın üzerine gelmiştir.</p>	<p>32. The blue pole of 20 meters has been broken due to storm and the tip of the pole is touching the top of the wall 8 meters away.</p>
	 <p>So, what is the height of the wall?</p>
<p>Buna göre, duvarın yüksekliği kaç metredir?</p> <p>A) 2 B) 3 C) 4 D) 5 E) 6</p>	<p>A) 2 B) 3 C) 4 D) 5 E) 6</p>
<p>34. Dikdörtgen şeklinde bir kâğıt; önce kısa kenarına paralel olan AB doğrusu boyunca Şekil 1'deki gibi ok yönünde, sonra uzun kenarına paralel olan CD doğrusu boyunca Şekil 2'deki gibi ok yönünde katlanarak Şekil 3 elde ediliyor.</p>	<p>34. Dikdörtgen şeklinde bir kâğıt; önce kısa kenarına paralel olan AB doğrusu boyunca Şekil 1'deki gibi ok yönünde, sonra uzun kenarına paralel olan CD doğrusu boyunca Şekil 2'deki gibi ok yönünde katlanarak Şekil 3 elde ediliyor.</p>
	
<p>Son şekilde oluşan dikdörtgenlerin alanları a, b, c ve d birimkaredir.</p>	<p>Son şekilde oluşan dikdörtgenlerin alanları a, b, c ve d birimkaredir.</p>
<p>Buna göre, başlangıçta kullanılan kâğıdın alanının a, b, c ve d türünden ifadesi aşağıdakilerden hangisidir?</p>	<p>Buna göre, başlangıçta kullanılan kâğıdın alanının a, b, c ve d türünden ifadesi aşağıdakilerden hangisidir?</p>
<p>A) $a + 2b + 3c + 4d$ B) $a + 2b + 2c + 2d$ C) $a + 2b + 2c + 3d$ D) $a + 2b + 4c + 2d$ E) $2a + 2b + 2c + 2d$</p>	<p>A) $a + 2b + 3c + 4d$ B) $a + 2b + 2c + 2d$ C) $a + 2b + 2c + 3d$ D) $a + 2b + 4c + 2d$ E) $2a + 2b + 2c + 2d$</p>

Table B.2, the MS items, which is identified with DIF in all methods, also represents as original (Turkish) and translated (English) languages. Original items were taken from ÖSYM website, whereas the items were translated a private translation office in Turkey.

Table B.2. *The MS Items, which is Identified with DIF in All Methods.*

Original Item	Translated Item
<p>1. a bir gerçel sayı olmak üzere, karmaşık sayılarda</p> $\frac{1-ai}{a-i} = i$ <p>eşitliği veriliyor.</p> <p>Buna göre, a kaçtır?</p> <p>A) 4 B) 3 C) 2 D) 1 E) 0</p>	<p>1. a is a real number. In complex numbers, the following equation is given.</p> $\frac{1-ai}{a-i} = i$ <p>What is a?</p> <p>A) 4 B) 3 C) 2 D) 1 E) 0</p>
<p>2. x, y ve z birbirinden farklı birer asal sayı olmak üzere,</p> $x(z-y) = 18$ $y(z-x) = 40$ <p>eşitlikleri veriliyor.</p> <p>Buna göre, x + y + z toplamı kaçtır?</p> <p>A) 17 B) 19 C) 21 D) 23 E) 25</p>	<p>2. x, y, and z are different prime numbers.</p> $x(z-y) = 18$ $y(z-x) = 40$ <p>According to the equation given above, what is x + y + z ?</p> <p>A) 17 B) 19 C) 21 D) 23 E) 25</p>
<p>3. n ve k pozitif tam sayılar olmak üzere, $\boxed{n_k}$ değeri</p> <ul style="list-style-type: none"> • n sayısı, k sayısına tam bölünüyorsa $\boxed{n_k} = \frac{n}{k}$ • n sayısı, k sayısına tam bölünmüyorsa $\boxed{n_k} = 0$ <p>olarak tanımlanıyor.</p> <p>Örnek:</p> $\boxed{10_2} = 5$ $\boxed{10_3} = 0$ <p>Buna göre,</p> $\boxed{n_2} + \boxed{n_3} = 10$ <p>eşitliğini sağlayan n sayılarının toplamı kaçtır?</p> <p>A) 24 B) 28 C) 32 D) 36 E) 40</p>	<p>3. n and k are positive integer numbers. $\boxed{n_k}$ is described as follows:</p> <ul style="list-style-type: none"> • if n is divisible by k, $\boxed{n_k} = \frac{n}{k}$ • if n is not divisible by k $\boxed{n_k} = 0$ <p>For example:</p> $\boxed{10_2} = 5$ $\boxed{10_3} = 0$ <p>So,</p> $\boxed{n_2} + \boxed{n_3} = 10$ <p>What is the sum of n values for the equation above?</p> <p>A) 24 B) 28 C) 32 D) 36 E) 40</p>

Table B.2. (Continued)

Original Item	Translated Item																																																
<p>5. a, b ve c sıfırdan farklı birer gerçel sayı olmak üzere,</p> $p : a + b = 0$ $q : a + c < 0$ $r : c < 0$ <p>önergeleri veriliyor.</p> $(p \wedge q) \Rightarrow r$ <p>önermesi yanlış olduğuna göre; a, b ve c sayılarının işaretleri sırasıyla aşağıdakilerden hangisidir?</p> <p>A) -, +, + B) -, +, - C) -, -, + D) +, -, + E) +, -, -</p>	<p>5. a, b, and c are non-zero real numbers. The following propositions are given.</p> $p : a + b = 0$ $q : a + c < 0$ $r : c < 0$ <p>If the proposition given below</p> $(p \wedge q) \Rightarrow r$ <p>is wrong, what are the symbols of a, b, and c, respectively?</p> <p>A) -, +, + B) -, +, - C) -, -, + D) +, -, + E) +, -, -</p>																																																
<p>6. a ve b tam sayılar olmak üzere, $a b$ gösterimi, a sayısının b sayısını tam böldüğünü ifade eder. Bir öğrenci,</p> <p>"a, b ve c tam sayıları $a c$ ve $b c$ koşullarını sağlıyorsa $(a + b) c$ koşulunu da sağlar."</p> <p>önermesinin yanlış olduğunu aksine örnek verme yöntemini kullanarak ispatlamak istiyor.</p> <p>Buna göre, öğrencinin verdiği örnek aşağıdakilerden hangisi olabilir?</p> <table border="1"> <thead> <tr> <th></th> <th>a</th> <th>b</th> <th>c</th> </tr> </thead> <tbody> <tr> <td>A)</td> <td>1</td> <td>3</td> <td>12</td> </tr> <tr> <td>B)</td> <td>2</td> <td>4</td> <td>24</td> </tr> <tr> <td>C)</td> <td>3</td> <td>2</td> <td>30</td> </tr> <tr> <td>D)</td> <td>4</td> <td>5</td> <td>60</td> </tr> <tr> <td>E)</td> <td>5</td> <td>1</td> <td>30</td> </tr> </tbody> </table>		a	b	c	A)	1	3	12	B)	2	4	24	C)	3	2	30	D)	4	5	60	E)	5	1	30	<p>6. a ve b tam sayılar olmak üzere, $a b$ gösterimi, a sayısının b sayısını tam böldüğünü ifade eder. Bir öğrenci,</p> <p>"a, b ve c tam sayıları $a c$ ve $b c$ koşullarını sağlıyorsa $(a + b) c$ koşulunu da sağlar."</p> <p>önermesinin yanlış olduğunu aksine örnek verme yöntemini kullanarak ispatlamak istiyor.</p> <p>Buna göre, öğrencinin verdiği örnek aşağıdakilerden hangisi olabilir?</p> <table border="1"> <thead> <tr> <th></th> <th>a</th> <th>b</th> <th>c</th> </tr> </thead> <tbody> <tr> <td>A)</td> <td>1</td> <td>3</td> <td>12</td> </tr> <tr> <td>B)</td> <td>2</td> <td>4</td> <td>24</td> </tr> <tr> <td>C)</td> <td>3</td> <td>2</td> <td>30</td> </tr> <tr> <td>D)</td> <td>4</td> <td>5</td> <td>60</td> </tr> <tr> <td>E)</td> <td>5</td> <td>1</td> <td>30</td> </tr> </tbody> </table>		a	b	c	A)	1	3	12	B)	2	4	24	C)	3	2	30	D)	4	5	60	E)	5	1	30
	a	b	c																																														
A)	1	3	12																																														
B)	2	4	24																																														
C)	3	2	30																																														
D)	4	5	60																																														
E)	5	1	30																																														
	a	b	c																																														
A)	1	3	12																																														
B)	2	4	24																																														
C)	3	2	30																																														
D)	4	5	60																																														
E)	5	1	30																																														
<p>7. a ve b sıfırdan farklı gerçel sayılar olmak üzere, gerçel sayılar kümesi üzerinde tanımlı bir f fonksiyonu</p> $f(ax + b) = x$ $f(a) = \frac{b}{a}$ <p>eşitliklerini sağlamaktadır.</p> <p>Buna göre, $f(0)$ değeri kaçtır?</p> <p>A) $\frac{-1}{2}$ B) $\frac{-1}{3}$ C) $\frac{-2}{3}$ D) 1 E) 2</p>	<p>7. a and b are non-zero real numbers. Equations for a function of f, defined in the set of real numbers, are given below.</p> $f(ax + b) = x$ $f(a) = \frac{b}{a}$ <p>So, what is $f(0)$?</p> <p>A) $\frac{-1}{2}$ B) $\frac{-1}{3}$ C) $\frac{-2}{3}$ D) 1 E) 2</p>																																																

Table B.2. (Continued)

Original Item	Translated Item
<p>9. Gerçek katsayılı ve baş katsayısı 1 olan 4. dereceden bir $P(x)$ polinomu her x gerçel sayısı için</p> $P(x) = P(-x)$ <p>eşitliğini sağlamaktadır.</p> $P(2) = P(3) = 0$ <p>olduğuna göre, $P(1)$ kaçtır?</p> <p>A) 12 B) 18 C) 24 D) 30 E) 36</p>	<p>9. A 4th degree $P(x)$ polynomial, with a leading coefficient of 1, of real multiples, the equation below</p> $P(x) = P(-x)$ <p>is true for every x real number.</p> $P(2) = P(3) = 0$ <p>Since the equation above is true as well, what is $P(1)$?</p> <p>A) 12 B) 18 C) 24 D) 30 E) 36</p>
<p>11. $\log_4 x$ ve $\log_8 \frac{1}{x}$ sayılarının aritmetik ortalaması $\frac{1}{2}$'dir.</p> <p>Buna göre, $\log_{16} x$ ifadesinin değeri kaçtır?</p> <p>A) $\frac{1}{2}$ B) $\frac{3}{2}$ C) $\frac{5}{2}$</p> <p>D) $\frac{1}{4}$ E) $\frac{5}{4}$</p>	<p>11. The arithmetic mean of $\log_4 x$ and $\log_8 \frac{1}{x}$ is $\frac{1}{2}$.</p> <p>So, what is $\log_{16} x$?</p> <p>A) $\frac{1}{2}$ B) $\frac{3}{2}$ C) $\frac{5}{2}$</p> <p>D) $\frac{1}{4}$ E) $\frac{5}{4}$</p>
<p>12. Terimleri birbirinden farklı ve ortak farkı r olan bir (a_n) aritmetik dizisi için</p> $a_1 = 3 \cdot r$ $a_6 = a_2 \cdot a_4$ <p>eşitlikleri veriliyor.</p> <p>Buna göre, a_{10} kaçtır?</p> <p>A) 10 B) 8 C) 6 D) 4 E) 2</p>	<p>12. For an arithmetic sequence of (a_n) that has different terms and a common difference of r, the following equation is given.</p> $a_1 = 3 \cdot r$ $a_6 = a_2 \cdot a_4$ <p>So, what is a_{10}?</p> <p>A) 10 B) 8 C) 6 D) 4 E) 2</p>

Table B.2. (Continued)

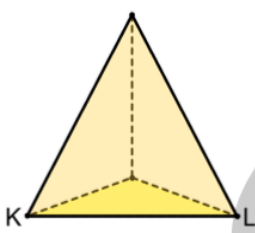
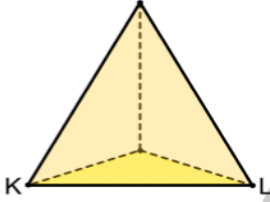
Original Item	Translated Item
<p>14. m ve n sıfırdan ve birbirinden farklı iki gerçel sayı olmak üzere,</p> $x^2 + (m+1)x + n - m = 0$ <p>denkleminin köklerinden biri $m - n$ sayıdır.</p> <p>Buna göre, $\frac{n}{m}$ oranı kaçtır?</p> <p>A) 2 B) 3 C) 4 D) 5 E) 6</p>	<p>14. m and n are two non-zero and different real numbers.</p> $x^2 + (m+1)x + n - m = 0$ <p>One of the roots of the equation above is number $m - n$</p> <p>So, what is $\frac{n}{m}$?</p> <p>A) 2 B) 3 C) 4 D) 5 E) 6</p>
<p>15. Bir sözcükte harflerin soldan sağa sıralanışıyla sağdan sola sıralanışı aynıysa bu sözcüğe bir palindrom sözcük denir.</p> <p>Örneğin; NEDEN, bir palindrom sözcüktür.</p> <p>Engin, birbirinden farklı 3 sesli ve 4 sessiz harfin her birini <u>istediği sayıda</u> kullanarak 5 harfli bir palindrom sözcük oluşturacaktır. Bu sözcükte iki sesli harfin yan yana gelmemesi ve iki sessiz harfin de yan yana gelmemesi gerekmektedir.</p> <p>Buna göre, Engin bu koşulları sağlayan kaç farklı palindrom sözcük oluşturabilir?</p> <p>A) 72 B) 84 C) 96 D) 108 E) 120</p>	<p>15. Palindrome is a word that is still the same when spelled backwards.</p> <p>For instance, the word MADAM is a palindrome.</p> <p>Engin is trying to produce a 5-letter palindrome using 3 vowels and 4 consonants. He can use every letter <u>as much as he wants</u>. In order to achieve this, he can't put two vowels or consonants together without any letter between them.</p> <p>So, under these terms, how many different palindromes can Engin produce?</p> <p>A) 72 B) 84 C) 96 D) 108 E) 120</p>
<p>16. Bir düzgün dörtüzlünün K ve L köşelerinde birer karınca bulunmaktadır.</p>  <p>Bu karıncalardan her biri buldukları köşelerden çıkan ayrıtlardan birini rastgele seçip bu ayrıtlar boyunca yürümeye başlıyor, ayrıtın diğer köşesine ulaştığında ise duruyor.</p> <p>Buna göre, karıncaların karşılaşma olasılığı kaçtır?</p> <p>A) $\frac{1}{3}$ B) $\frac{2}{3}$ C) $\frac{1}{4}$</p> <p>D) $\frac{3}{4}$ E) $\frac{1}{9}$</p>	<p>16. There is one ant on K and L corners of a regular tetrahedron.</p>  <p>Each of these ants select one edge and start walking along that edge. When they reach to the end of that specific edge, they stop.</p> <p>So, what is the probability of those two ants meeting?</p> <p>A) $\frac{1}{3}$ B) $\frac{2}{3}$ C) $\frac{1}{4}$</p> <p>D) $\frac{3}{4}$ E) $\frac{1}{9}$</p>

Table B.2. (Continued)

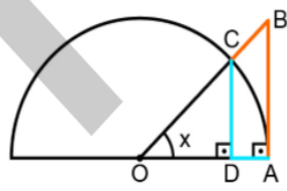

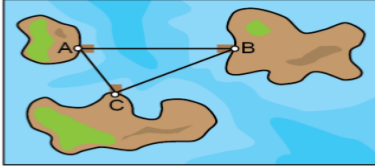
Original Item	Translated Item
<p>18. $f(x) = \begin{cases} 10 - x^2 & , \quad x < 0 \\ ax + b & , \quad 0 \leq x \leq 3 \\ (1 - x)^2 & , \quad x > 3 \end{cases}$</p> <p>fonksiyonu gerçel sayılar kümesi üzerinde süreklidir.</p> <p>Buna göre, $a + b$ toplamı kaçtır?</p> <p>A) 16 B) 15 C) 12 D) 9 E) 8</p>	<p>18. $f(x) = \begin{cases} 10 - x^2 & , \quad x < 0 \\ ax + b & , \quad 0 \leq x \leq 3 \\ (1 - x)^2 & , \quad x > 3 \end{cases}$</p> <p>The function above is defined in the set of real numbers and continuous.</p> <p>So, what is $a + b$?</p> <p>A) 16 B) 15 C) 12 D) 9 E) 8</p>
<p>22. Bir internet şirketi en fazla 1000 müşteriye hizmet verebilmekte ve aylık internet ücretini 40 TL olarak belirlediğinde bu sayıya ulaşabilmektedir. Bu şirket aylık internet ücretinde yaptığı her 5 TL'lik artış sonrasında müşteri sayısında 50 azalma olduğunu gözlemlemiştir.</p> <p>Bu şirket, aylık internet ücretinden elde edeceği toplam gelirin en fazla olması için aylık internet ücretini kaç TL olarak belirlemelidir?</p> <p>A) 55 B) 60 C) 65 D) 70 E) 75</p>	<p>22. An internet service provider company is capable of provide service to a maximum of 1000 customers and can reach that number by setting the monthly price to 40 TL. Every 5 TL increase in the price results in 50 decrease in the number of customers.</p> <p>In order to maximize their profit, what should this company's monthly price be?</p> <p>A) 55 B) 60 C) 65 D) 70 E) 75</p>
<p>29. Aşağıda, O merkezli yarıçapı 1 birim olan yarım çember ile OAB ve ODC dik üçgenleri gösterilmiştir. A ve C noktaları hem OAB üçgeninin hem de yarım çemberin üzerindedir.</p>  <p>Buna göre,</p> $\frac{ AB + BC }{ CD + DA }$ <p>oranının x türünden eşiti aşağıdakilerden hangisidir?</p> <p>A) $\sin x$ B) $\tan x$ C) $\cot x$</p> <p>D) $\csc x$ E) $\sec x$</p>	<p>29. There is a semicircle with O as the center and 1 unit radius, and OAB and ODC right angled triangles. A and C are on both OAB triangle and the semicircle.</p>  <p>So,</p> $\frac{ AB + BC }{ CD + DA }$ <p>What is the x equivalent of the equation above?</p> <p>A) $\sin x$ B) $\tan x$ C) $\cot x$</p> <p>D) $\csc x$ E) $\sec x$</p>

Table B.2. (Continued)

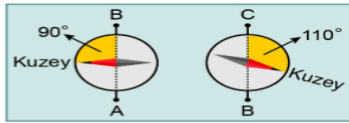
Original Item

Translated Item

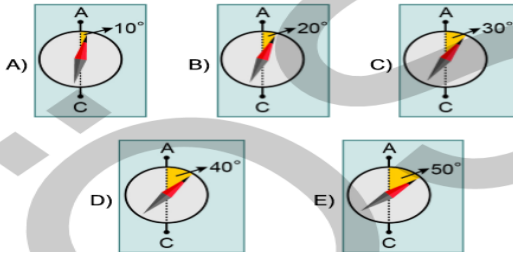
30. Temel Kaptan, teknesindeki turistleri sabah A adasından B adasına, öğlen B adasından C adasına, akşam da C adasından A adasına götürecektir. Teknenin adalardaki iskelelerde duracağı noktalar, AB kenarı BC kenarına eşit olan bir ABC üçgeninin köşe noktaları olarak şekildeki gibi işaretlenmiştir.



Temel Kaptan dönüş yolunda karanlıkta seyahat edeceğini bildiğinden A'dan B'ye ve B'den C'ye ilerlerken pusulasının kuzeyi gösteren ibresi ile izlediği yol arasındaki açıyı bir kâğıda aşağıdaki gibi not almıştır.



Buna göre, Temel Kaptan C'den A'ya gitmek için pusulasını aşağıdakilerden hangisi gibi ayarlamalıdır?

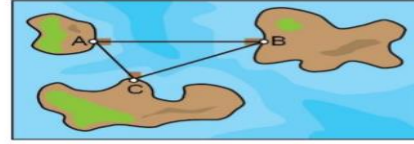


31. Dik koordinat düzleminde; bir köşesi orijinde, diğer köşeleri ise $y = x$ ve $y = -x$ doğruları üzerinde olan bir üçgenin kenarortayları $(2, 4)$ noktasında kesişmektedir.

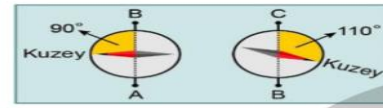
Buna göre, bu üçgenin alanı kaç birimkaredir?

- A) 18 B) 24 C) 27
D) $9\sqrt{2}$ E) $18\sqrt{2}$

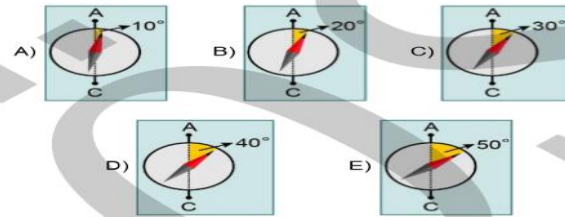
- Captain Temel is taking the tourists from Island A to Island B in the morning, from Island B to Island C at noon, and from Island C to Island A in the evening. The points on the piers at which the boat will stop are illustrated below as the corners of an ABC triangle in which AB and BC are equal.



Since Captain Temel knows that they are going to be traveling in the dark on the way back, he draws the angle between the north indicator of his compass and the direction they are going towards on a piece of paper while going from Island A to Island B and Island B to Island C.



So, how should Captain Temel set his compass to go from Island C to Island A?



31. On the cartesian coordinate plane, the medians of a triangle with one corner on the origin and other corners on $y = x$ and $y = -x$ are crossing at $(2, 4)$.

So, what is the area of the triangle in units square?

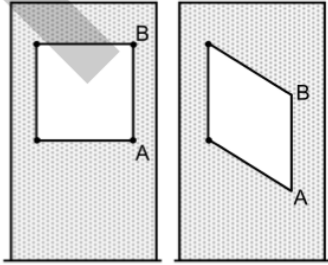
- A) 18 B) 24 C) 27
D) $9\sqrt{2}$ E) $18\sqrt{2}$

Table B.2. (Continued)

Original Item

Translated Item

32. Eşit uzunlukta dört telin birbirine monte edilmesiyle oluşturulan ve Şekil 1'deki gibi çivilerle köşelerinden duvara sabitlenen kare biçiminde bir çerçevenin duvarda kapladığı alan 100 birimkaredir.



Şekil 1

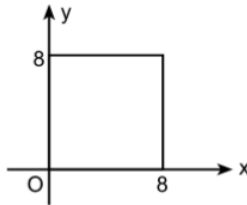
Şekil 2

A ve B köşeleri üzerindeki çivilerin çıkması sonucu bir tarafının aşağı kaymasıyla Şekil 2'deki gibi bir eşkenar dörtgen hâlini alan bu çerçevede A ve B köşelerinin yerden yüksekliği 6'şar birim azalmış, diğer iki köşenin konumu ise değişmemiştir.

Buna göre, çerçevenin duvarda kapladığı alan kaç birimkare azalmıştır?

- A) 18 B) 20 C) 26 D) 30 E) 32

33.



Dik koordinat düzleminde verilen şekildeki kare, eğimi $-\frac{1}{4}$ olan bir doğru ile eşit alanlı iki bölgeye ayrılıyor.

Bu doğru x-eksenini $(a, 0)$ noktasında kestiğine göre, a kaçtır?

- A) 12 B) 14 C) 16 D) 18 E) 20

34. Dik koordinat düzleminde birinin merkezi $(12, 0)$ noktası, diğ erinin merkezi ise $(0, 9)$ noktası olan iki çember sadece $(4, 6)$ noktasında kesişmektedir.

Bu çemberlerin orijine en yakın olan noktaları arasındaki uzaklık kaç birimdir?

- A) $\sqrt{5}$ B) $\sqrt{10}$ C) $\sqrt{13}$
D) $2\sqrt{5}$ E) $2\sqrt{10}$

32. The area of the frame in the shape of a square of four wires mounted together on the wall as in Figure 1 is 100 unit square.

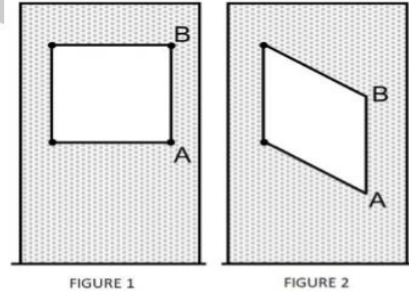


FIGURE 1

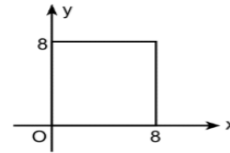
FIGURE 2

The frame has taken the shape of a rhombus, as in Figure 2, as a result of the drop of A and B, and the height of those points have been lowered 6 units each. Other points haven't moved at all.

So, how many units square did the area of the frame decrease?

- A) 18 B) 20 C) 26 D) 30 E) 32

33. The square given below is divided into two equal pieces by a line with a slope of $-\frac{1}{4}$.



So, if that line crosses the x axis at $(a, 0)$, what is a?

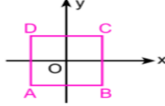
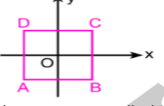
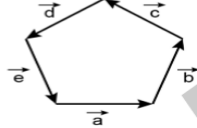
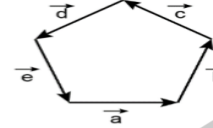
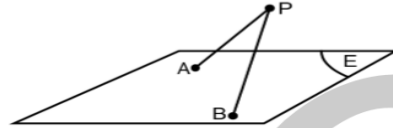
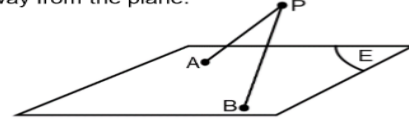
- A) 12 B) 14 C) 16 D) 18 E) 20

34. On the cartesian coordinate plane, two circle, whose centers are $(12, 0)$ and $(0, 9)$, are crossing each other only at $(4, 6)$

What is the distance between the closest points of those circles?

- A) $\sqrt{5}$ B) $\sqrt{10}$ C) $\sqrt{13}$
D) $2\sqrt{5}$ E) $2\sqrt{10}$

Table B.2. (Continued)

Original Item	Translated Item
<p>35. Dik koordinat düzleminde köşe noktalarının koordinatları $A(-1, -1)$, $B(1, -1)$, $C(1, 1)$, $D(-1, 1)$ olan ABCD karesi aşağıda verilmiştir.</p>  <p>Bu kareye sırasıyla</p> <ul style="list-style-type: none"> • orijin etrafında saat yönünün tersine 45° döndürme, • y-eksenine göre yansım, • orijin etrafında saat yönünde 45° döndürme dönüşümleri uygulanıyor. <p>Son durumda bu karenin koordinatları değişmeyen köşe noktaları aşağıdakilerden hangisidir?</p> <p>A) A ve B B) A ve C C) A ve D D) B ve C E) C ve D</p>	<p>35. On the cartesian coordinate plane, the square with the corners of $A(-1, -1)$, $B(1, -1)$, $C(1, 1)$, $D(-1, 1)$ is given below.</p>  <p>The following changes are applied to the square respectively.</p> <ul style="list-style-type: none"> • 45° counter clock wise turn around the origin. • Reflection against y axis. • 45° clock wise turn around the origin <p>So, which corners of the square have stayed the same?</p> <p>A) A and B B) A and C C) A and D D) B and C E) C and D</p>
<p>37. Analitik düzlemde verilen bir düzgün beşgenin kenarları şekildeki gibi \vec{a}, \vec{b}, \vec{c}, \vec{d} ve \vec{e} vektörleri olarak adlandırılmıştır.</p>  <p>Buna göre, bu beş vektör arasından rastgele seçilen iki vektörün iç çarpımının pozitif olma olasılığı kaçtır?</p> <p>A) $\frac{1}{2}$ B) $\frac{1}{5}$ C) $\frac{2}{5}$ D) $\frac{1}{10}$ E) $\frac{3}{10}$</p>	<p>37. The edges of a pentagon on analytical plane is represented by vectors of \vec{a}, \vec{b}, \vec{c}, \vec{d}, and \vec{e}.</p>  <p>So, what is the probability of two random vectors' inner product being positive?</p> <p>A) $\frac{1}{2}$ B) $\frac{1}{5}$ C) $\frac{2}{5}$ D) $\frac{1}{10}$ E) $\frac{3}{10}$</p>
<p>39. Ayrıt uzunluğu 1 birim olan 3 adet küp, her birinin en az bir yüzü diğer bir küpün bir yüzüyle tam örtüşecek biçimde birbirine yapıştırılıyor.</p> <p>Buna göre, bu şekilde elde edilebilecek bir cismin seçilen iki köşesi arasındaki uzaklık birim türünden aşağıdakilerden hangisi <u>olamaz</u>?</p> <p>A) $\sqrt{7}$ B) $\sqrt{8}$ C) $\sqrt{9}$ D) $\sqrt{10}$ E) $\sqrt{11}$</p>	<p>39. 3 cubes, with edge length of 1 unit, are glued together to cover one side of one cube completely covering one side of another cube.</p> <p>So, which of the below can't be the distance between two corners of such shape in units?</p> <p>A) $\sqrt{7}$ B) $\sqrt{8}$ C) $\sqrt{9}$ D) $\sqrt{10}$ E) $\sqrt{11}$</p>
<p>40. Uzayda bir E düzlemi üzerinde A ve B noktaları ve bu düzleme 4 birim uzaklıkta bir P noktası veriliyor.</p>  <p>PA ve PB doğru parçalarının E düzlemi üzerine dik izdüşümleri ile AB doğru parçası, kenar uzunluğu 2 birim olan bir eşkenar üçgen oluşturmaktadır.</p> <p>Buna göre, $PA \cdot PB$ çarpımı kaçtır?</p> <p>A) 8 B) 12 C) 16 D) 18 E) 20</p>	<p>40. A and B are on plane E. P is on the space and 4 unit away from the plane.</p>  <p>With the projections of PA and PB line segments on the plane E, AB line segment creates an equilateral triangle with edges of 2 units.</p> <p>So, what is $PA \cdot PB$?</p> <p>A) 8 B) 12 C) 16 D) 18 E) 20</p>