ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(3); 283-301

Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions *

Mustafa ÇELİK ** Yeşim ÖZER ÖZKAN ***

Abstract
It is accepted that the presence of Differential Item Functioning (DIF) in large scale examinations may be an indication of bias. The aim of the present study was to analyze whether dichotomous (1-0) items in the PISA 2015 mathematics subtest exhibit DIF with regard to gender and statistical regions. The study was carried out using the data of 2409 students who took part in PISA 2015 and answered all mathematics questions. Rasch model was used via Winsteps software to determine whether the items exhibit DIF or not. Confirmatory Factor Analysis (CFA) was applied to the sixty-three mathematics questions in the clusters. The modification indices and goodness of fit values were examined for CFA and a total of eight items that disrupt the model structure were excluded from the test. DIF analyses were carried out for the 55 items that were observed to be one-dimensional. The analysis results based on gender indicated that five items exhibit DIF. Two of these items exhibit DIF in favor of girls, while three in favor of men. Statistically significant DIF findings were observed in all items when the analyses results based on statistical regional units were analyzed. While at least 10 DIF cases were observed as a result of the binary territory comparison on an item basis, maximum 38 DIF cases were observed. Minimum DIF was observed in item Q4 as a result of regional comparisons, whereas maximum DIF was observed in items Q47 and Q50.

Key Words: Differential item functioning, Rasch model, bias, statistical regions, PISA.

INTRODUCTION

Structuring of labor via qualified education is directly related to education quality and policy. Continuous advancement of scientific developments by way of innovations increases the importance of cooperative education quality and education policies. Measurement and evaluation tools are used for determining the personal qualifications of individuals who undergo a certain education. Individual outputs determined subject to the implementation purpose provide insight into the competence of the individuals (Baykul, 2000). Large scale examinations are conducted for the global evaluation of educated individuals. Large scale examinations in Turkey are conducted by the Ministry of National Education (MoNE), General Directorate of Measurement, Evaluation and Examination Services and Student Selection and Placement Center.

Large scale examinations are carried out worldwide such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS) are carried out which enable the comparison of many education-related outputs.

PISA, organized by the Organization for Economic Cooperation and Development (OECD) has been applied in Turkey once every three years since 2000. The fundamental knowledge and skills of the students in the fields of science, mathematics and reading are assessed as part of the PISA project for a 15-year-old group of students (MoNE, 2013). PISA goes beyond assessing whether students can

* This study was a part of thesis was produced from the first author's master thesis.
** Teacher, Ministry of National Education, Gaziantep-Turkey, mstfclk47@gmail.com, ORCID ID: 0000-0002-9501-8726
*** Assoc. Prof., Gaziantep University, Faculty of Education, Gaziantep-Turkey, e-posta:yozer80@gmail.com, ORCID ID:0000-0002-7712-658X

To cite this article:
Çelik, M. & Özer-Özkan, Y.(2020). *Analysis of Differential Item Functioning of PISA 2015 Mathematics Subtest Subject to Gender and Statistical Regions, 11*(3), 283-301. doi: 10.21031/epod.715020.

Received: 05.04.2020
Accepted: 10.07.2020

reproduce what they have learned in school, thus focusing on determining their ability to apply their knowledge in real life, solve problems in novel situations as well as their abilities to make use of skills such as extrapolation and reasoning (MoNE, 2010). The data acquired from these surveys contribute to the interpretation of cognitive data. Turkey has taken part in school and student surveys within the scope of PISA 2015 (MoNE, 2015).

Individuals who take part in international large scale examinations vary with regard to characteristics such as ethnic origins, language, culture, etc. The balance of presence in life of female-male individuals differs among different societies, especially within the context of gender. These differences make it difficult to adapt the tests into different languages and cultures in intercultural studies (Van de Vijver & Tanzer, 2004). Large scale examinations should be prepared without allowing for any inequalities by taking the aforementioned circumstances under control. The prepared test items should not provide any advantage or disadvantage to any group (Öğretmen, 1995).

The validity and reliability of the test scores of the individuals for the characteristics to be compared with regard to desired ability or success may have an impact on the accuracy of the decisions made based on these data (Gierl, 2000). The fact that large scale examinations used as a resource for important decisions are free of errors brings forth the validity of the test or test items and thus, the presence of bias observed as systematic error. Item and test bias is among the threats against meeting the validity requirement (Clauser & Mazor, 1998). In this regard, unbiasedness is considered as a criterion in order for a test or test item to meet the validity requirement (Camili & Shepard, 1994). Characteristics of measurement such as the ability level, item discrimination, item difficulty, distribution, reliability vary subject to the group (Özer Özkan, 2012). In this regard, it is expected that the psychometric characteristics of the measurement tool are the same for all responders (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017).

Bias is accepted as the presence of systematic error in the test items (Osterlind, 1983). It can be indicated that the value of the variables is systematically low or high in case of bias in the test item (Çıkrıkçı Demirtaşlı & Uluştaş, 2015). It is expected that the individuals will have the same probability of responding to the items correctly if the test or test items have the same construct validity for all individuals in the group (Camili & Shepard, 1994). Determining the test or test item that exhibits DIF is important for validity. In this case, the validity of the study carried out will be at risk in case studies are not carried out for determining the biased items in large scale tests.

It is observed that studies on DIF are carried out frequently in different countries in order to test the validity of international large scale examinations. DIF and bias studies based on gender, culture and language are observed frequently in literature. It has been observed that DIF and bias studies have been carried out for international or national scale examinations subject to *gender* (Acar, 2011; Alkaline, 2014; Amour, AL-Gadarene Alomar, & AL Ruairi, 2015; Arıffin, Idris, & Ihsak, 2010; Ateşok Deveci, 2008; Bakan Kalaycıoğlu, 2008; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Bekci, 2007; Berberoğlu, 1995; Birjandi & Amini, 2007; Doolite & Cleary, 1987; Gamer & Engelhard, 1999; Hanna, 1986; Haris & Cartlon, 1993; Karakaya, 2012; Karakaya & Cult, 2012; Ken, Sunbelt, & Omar, 2013;Kıbrıslıoğlu & Atalay Kabasakal, 2017; Kurnaz, 2006;Latifi, Bulut, Gierl, Christie, & Jeeva, 2016; Le, 1999,2009; Lyons-Thomas, Sandilands, & Ercikan, 2014; Öğretmen & Doğan, 2004; Özer Özkan & Fincan, 2017; Satıcı & Özer Özkan, 2016; Sunna, 2012;Şenferah, 2015; Taylor & Lee, 2012;Turkman, 2014;Ultras, 2012;Yilin & Tavşancıl, 2015;Yurdugül & Aşkar, 2004; Zenisky, Hambleton, & Robin, 2004; Zwick & Ercikan, 1989); *school type* (Bakan Kalaycıoğlu, 2008; Bekci, 2007; Karakaya & Kutlu, 2012; Şenferah, 2015); *regions and cultures* (Ercikan & Kim, 2009; Gök, Atalay Kabasakal, & Kelecioğlu, 2014; Özmen, 2014; Ulutaş, 2012; Yurdugül & Aşkar, 2004; Zwick & Ercikan, 1989).

It is observed that bias studies in Turkey have been carried out since the early 1990s on data for national tests of Secondary School Institutions Student Selection and Placement Test, Student Selection Test, Level Determination Exam and Transition from Primary to Secondary Education. There has been an increase in DIF and bias studies on PISA, TIMSS and PIRLS following the increase in the popularity of international large scale examinations after the 2000s. Bias studies have been included in the PISA 2015 technical report published by the OECD and information has been presented on how the biased items are controlled (OECD, 2015). It is observed that there is no mention of any DIF or bias towards gender

and statistical regions in both the PISA 2015 technical report published by OECD and the PISA 2015 National Report published by MoNE General Directorate of Measurement, Evaluation and Examination Services and Student Selection. Moreover, the fact that the impacts of gender and culture yield results that cannot be ignored in DIF and bias studies in literature increase the importance of the study. In this regard, it is expected that the present study aiming to examine the DIF of PISA 2015 mathematics subtest subject to gender and statistical regions will contribute to the related studies in the literature.

The Aim of the Study

The aim of the study is to determine whether the dichotomous (1-0) items in the PISA 2015 application mathematics subtest exhibit DIF subject to gender and territory or not, according to the Item Response Theory (IRT) via the Rasch Model. Answers to the following questions were sought within the framework of this aim:

1. Do the binary scored items in the PISA 2015 mathematics literacy subtest exhibit DIF based on the analyses via the Rasch method?
2. Do the binary scored items in the PISA 2015 mathematics literacy subtest exhibit DIF with regard to the statistical regions in Turkey based on the analyses via the Rasch method?

METHOD

The study examines whether the dichotomous (1-0) items in the PISA 2015 mathematics literacy subtest exhibit DIF or not with regard to gender and regions. The study is designed based on considering the current situation from different perspectives, defining and comparing the relations between them and expressing them in a holistic and circumspect way (Büyüköztürk et al., 2013). In this regard, this is a descriptive study in the survey model.

Population and Sample

The 15-year-old student population who can take part in the PISA 2015 Turkey application was determined as 925.366. A total of 187 schools from 61 provinces representing the 12 regions in the Turkey Nomenclature of Territorial Units for Statistics. (NUTS) 1st level took part in the PISA 2015 application. Turkey's NUTS classifications are officially termed statistical regions. Therefore, in this study, the term statistical region is used. School sample groups were determined during the PISA 2015 study via a stratified random sampling method, while the students to take part in the application were selected via random sampling method (MoNE, 2015). The present study was carried out using the data of 2409 students who responded to the PISA 2015 mathematics subtest. The regionals DIF findings of the items in the PISA 2015 mathematics subtest were examined based on the NUTS-1 classification presented in Table 1. Table 1 presents the distribution subject to gender and statistical regions for the students who took the mathematics subtest in 2015.

Table 1. Sample Group Distribution for the Students Subject to Gender and Statistical Regions

	Female		Male		Total	
Territory	f	%	f	%	f	%
İstanbul	221	18.45%	210	17.34%	431	17.89%
Western Marmara	53	4.42%	46	3.80%	99	4.11%
Aegean	154	12.85%	130	10.73%	284	11.79%
Eastern Marmara	102	8.51%	107	8.84%	209	8.68%
Western Anatolia	112	9.35%	113	9.33%	225	9.34%
Mediterranean	160	13.36%	182	15.03%	342	14.20%
Central Anatolia	61	5.09%	70	5.78%	131	5.44%
Western Black Sea	54	4.51%	77	6.36%	131	5.44%
Eastern Black Sea	35	2.92%	45	3.72%	80	3.32%
Northeastern Anatolia	44	3.67%	36	2.97%	80	3.32%
Central Eastern Anatolia	60	5.01%	59	4.87%	119	4.94%
Southeastern Anatolia	142	11.85%	136	11.23%	278	11.54%
Total	1198	100.00%	1211	100.00%	2409	100.00%

It can be observed when Table 1 is examined that the students who responded to the PISA 2015 mathematics subtest have similar gender distributions. Whereas, it can be indicated that the distribution according to the statistical regions of the students who responded to the PISA 2015 is similar within the scope of NUTS.

Data Collection Tools

Procedure

PISA 2015 cognitive test results published at the OECD official website were used in the study. PISA 2015 was applied in Turkey by way of a computer-based assessment method instead of as a pencil-paper test. The items in the mathematics subtest were included in 36 of the 66 booklets prepared for the implementation of this method. PISA 2015 cognitive test data were downloaded after which the data related to the Turkey mathematics subtest were sorted out. The data for 2409 Turkish students who responded to all of the items in the mathematics literacy subtest were used (OECD, 2015).

Data Analysis

The items in PISA 2015 mathematics subtest were classified by the OECD into 6 + 1 (equivalent form) different clusters comprised of 11 or 12 questions. Each booklet used for implementation in PISA 2015 was prepared by including the question group that makes up one or two mathematics clusters. Annex 1 presents the additional data regarding the mathematics subtest items included in each booklet implemented in PISA 2015.

The forms used in PISA 2015 implementation were prepared consecutively, placing to the booklets each of the six different clusters. The students answered the items in the mathematics cluster in the booklet. Each student did not take the complete mathematics subtest. Instead, they answered the mathematics items in one or two of the six mathematics clusters determined by the OECD. The following sample distribution method was used for selecting the items in the forms prepared for PISA 2015 implementation. Table 2 presents the sample distribution method for the number of items included in the mathematics subtest clusters.

Table 2. Mathematics Subtest Items Sample Cluster Distribution

	Cluster01	Cluster02	Cluster03	Cluster04	Cluster05	Cluster06a
Forms	11 Items	12 Items	12 Items	11 Items	12 Items	11 Items
Form33	x	x				
Form34		x	x			
Form35			x	x		
Form36				x	x	
Form37					x	x
Form38	x					x
Form39		x			x	

While the items in Cluster01 and Cluster02 in Form33 were observed when Table 2 is examined, the items in Cluster05 and Cluster06 were answered in Form37. Two equivalent forms were prepared for the final cluster in the PISA 2015 mathematics subtest. One of the equivalent forms was applied in Turkey. The equivalent form was named as "Cluster06a" throughout the study. The data regarding the Turkey mathematics subtest declared by the OECD were extracted by taking the data related to gender, territory, form number and question responses. Six multi-scored items were excluded from the data cluster in addition to one item excluded by the OECD since the study was going to be carried out using the dichotomous (1-0) items.

Six copies of the data file were prepared since the responses to each of the mathematics subtest items included in each cluster will be analyzed separately. Each data file was renamed as such (e.g.; Cluster01, Cluster02, etc.), after which they were cleaned up so as to include only the responses to the items in that cluster. Hence, each file was prepared to include gender, education territory and the responses to the items in that cluster. One copy of each cluster file was made in order to ensure that the data are in compliance with the Winsteps 3.80.1 software. Microsoft Office Excel software was used to arrange one of the files to include gender and item responses and the other file to include regions and item responses. As a result, 12 data files were prepared by sorting the PISA 2015 mathematics subtest items for analysis in six clusters and two variables. Winsteps was run and each Excel file was transformed into a text document for analysis. The proper syntax was written for the file transformed into a text document in accordance with the Winsteps software. The related package software was used in the study for the analysis of the assumptions of IRT and for data sorting. The normality graph and the skewness-kurtosis coefficients were examined for the normal distribution assumption of the data of each cluster. The skewness and kurtosis coefficients of each cluster were determined to be between +1 and -1. CFA was applied on the PISA 2015 mathematics subtest for the unidimensionality analysis on a cluster basis. CFA was used to examine RMSEA, GFI, NFI, RFI, IFI, AGFI, CFI and SRMR with regard to validity and goodness of fit values. Two items in Cluster02, three items in Cluster04, one item in Cluster05 and two items in Cluster06a were excluded from the study since they did not meet the unidimensionality assumption. It was observed that the items of each cluster included in the study meet the normality, unidimensionality and local independence assumptions. Finally, it was accepted that the items in the PISA 2015 mathematics subtest are structured in accordance with the Rasch model, according to IRT.

RESULTS

Gender-Related DIF Findings for the Mathematics Subtest Items

This section examines the gender-related DIF values for the dichotomous (1-0) items included in six clusters in PISA 2015 mathematics subtest. DIF measurement value regarding the responses of female and male students to the nine items in Cluster01 of the mathematics subtest subject to the item codes,

the contrast value between the DIF measurement values and t values were examined with the findings presented in Table 3.

Table 3. DIF Values of Items in Cluster01 Subject to Gender

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q1		-1.94		-1.82	-.12	-.63
Q2		-1.12		-1.37	.26	1.41
Q3		-1.03		-1.34	.31	1.69
Q4		-.33		-.24	-.09	-.45
Q5	Female	.37	Male	.56	-.19	-.88
Q6		.11		.19	-.09	-.43
Q7		2.25		1.64	.60	1.91
Q8		1.13		1.76	-.62	-2.35
Q9		.59		.71	-.12	-.56

It was observed when Table 3 was examined that item Q7 exhibits DIF (.60) in favor of males, whereas item Q8 exhibits DIF (-.62) in favor of females. Figure 1presents the change in DIF for the items in Cluster01 subject to gender.

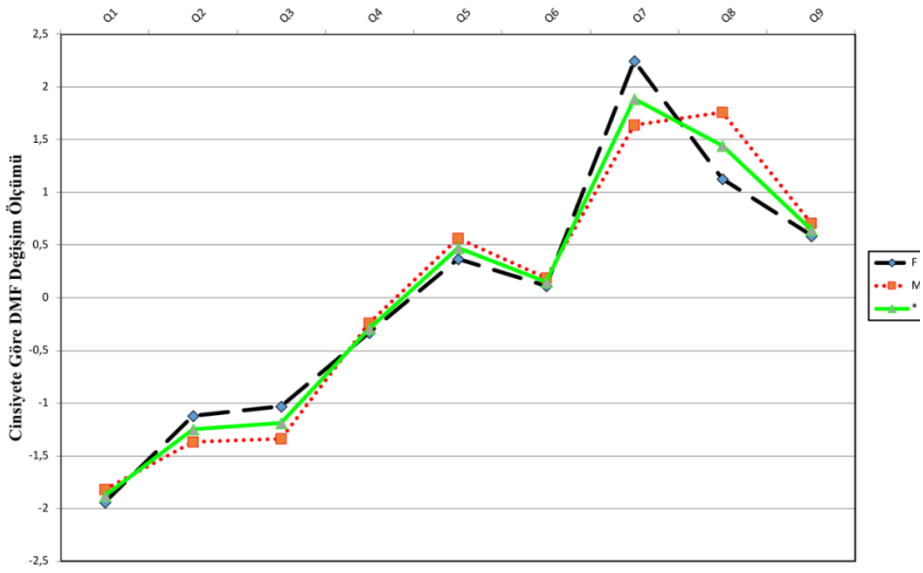

Figure 1. DIF Change Graph Subject to Gender for the Nine Items in Cluster01

It can be understood from the DIF change graph in Figure 1 that the items with contrast values outside the range of .5 and -.5 logit exhibit DIF at a statistically significant level. It can be observed that items Q7 and Q8 have the highest divergence from the mean value for the female and male students. In conclusion, it can be indicated when the DIF values of the nine items in Cluster01 subject to gender are examined that items Q7 and Q8 exhibit DIF at a statistically significant level. Table 4 presents the DIF value subject to gender for the items in Cluster02 of the mathematics subtest.

Table 4. DIF Values Subject to Gender for the Items in Cluster02

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q10	Female	-.03	Male	.44	-.47	-2.49
Q11		.83		.68	.15	.74
Q12		-.64		-.87	.22	1.25
Q13		-2.80		-2.94	.15	.62
Q14		2.70		2.05	.66	2.11
Q16		.31		.64	-.33	-1.71
Q18		-.31		.06	-.36	-2.01
Q19		.21		-.22	.43	2.35

It is observed in Table 4 that the item Q14 in Cluster02 (.66) exhibits DIF in favor of males. Figure 2 presents the DIF change graph of items in Cluster02 subject to gender.

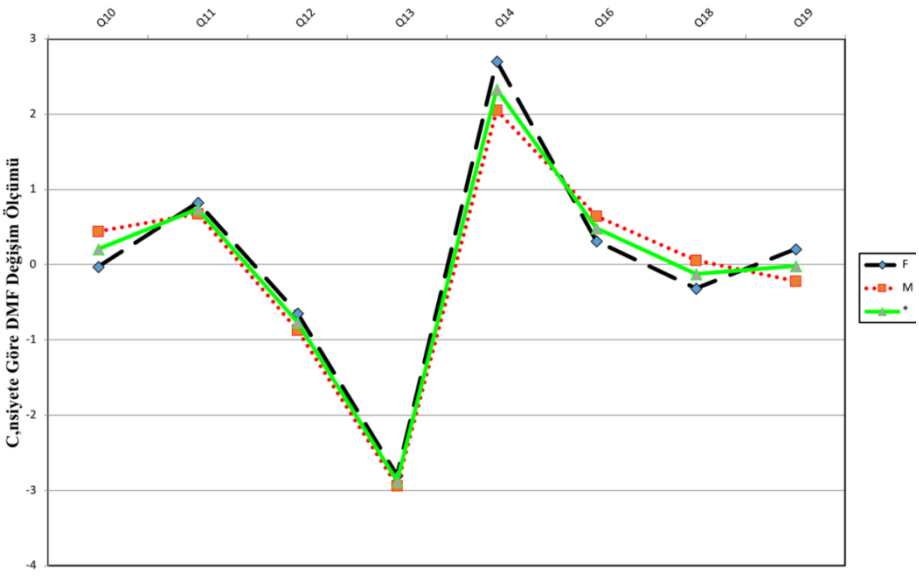

Figure 2. DIF Change Graph of the Eight Items in Cluster02 Subject to Gender

It can be stated when the DIF values of the eight items in Cluster02 are examined that the Q14 coded item exhibits DIF at a statistically significant level. Table 5 presents the DIF value of the items in the mathematics subtest Cluster03 subject to gender.

Table 5. DIF Values Subject to Gender for the Items in Cluster03

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t value
Q20		-.80		-.85	.05	.26
Q21		.37		-.09	.46	2.43
Q22		.29		.13	.17	.87
Q23		.35		.35	.00	.00
Q24		-1.62		-1.36	-.26	-1.39
Q25	Female	4.07	Male	4.56	-.49	-.80
Q26		-.65		-.78	.14	.74
Q27		-.77		-.59	-.18	-.97
Q28		1.15		1.25	-.10	-.44
Q29		1.18		1.56	-.38	-1.64
Q30		-3.86		-3.86	.00	.00

The DIF contrast values of the 11 items in Cluster03 have been calculated between 0.5 and -0.5 logit in Table 5. Thus, it can be stated that the items in Cluster03 do not work for or against any group. Figure 3 presents the change in DIF graph for the items in Cluster03 subject to gender.

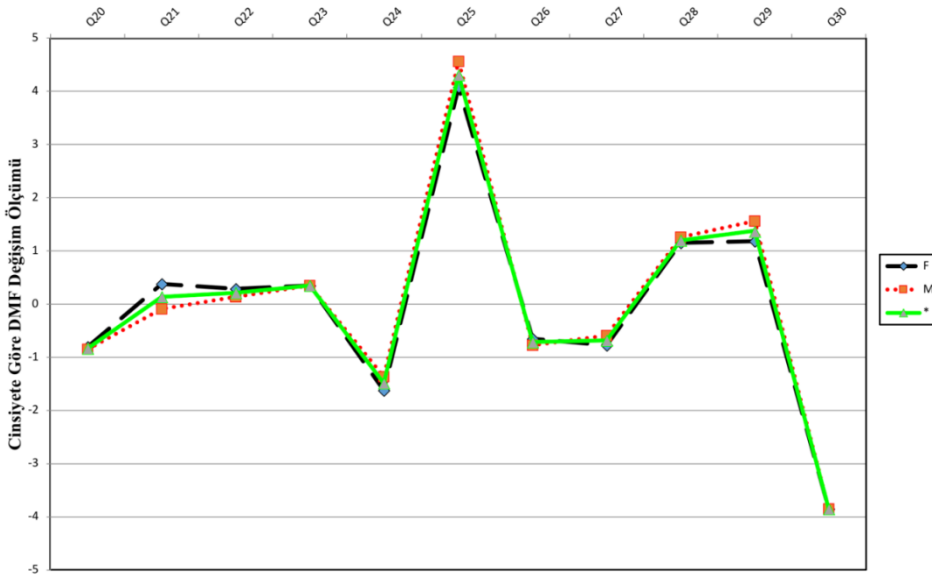

Figure 3. DIF Change Graph of the 11 Items in Cluster03 Subject to Gender

It can be stated when the change in DIF subject to the gender of the 11 items in Cluster03 presented in Figure 3 is examined that there are no items that exhibit DIF at a statistically significant level. Table 6 presents the DIF values subject to the gender of the items in Cluster04 of the mathematics subtest.

Table 6. DIF Values of the Items in Cluster04 Subject to Gender

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q31		-2.00		-2.00	.00	.00
Q33		-1.29		-1.18	-.11	-.63
Q34		.34		.54	-.20	-1.04
Q35	Female	-.89	Male	-1.08	.19	1.09
Q37		4.34		4.67	-.34	-.51
Q38		.01		-.07	.08	.43
Q39		-.74		-.68	-.06	-.37
Q40		.06		-.07	.13	.70

It can be stated when the DIF contrast values subject to the gender of the eight items in Cluster04 presented in Table 6 are examined that the DIF contrast value calculated between 0.5 and -0.5 logit and that the items do not exhibit DIF subject to gender.

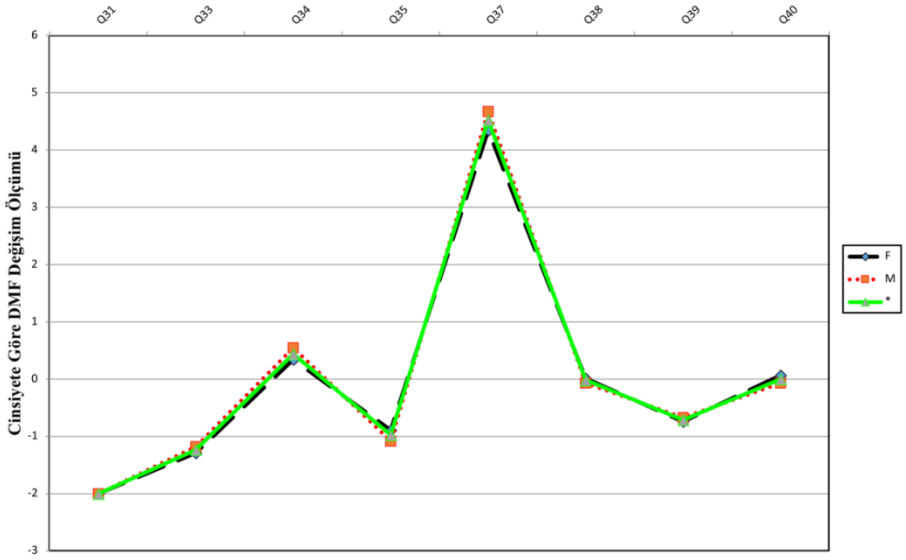

Figure 4. DIF Change Graph Subject to Gender for the Eight Items in Cluster04

It can be understood when Figure 4 is examined that there are no items with contrast values outside the boundaries of .5 and -.5 logit, or in other words, that the items in Cluster04 do not exhibit DIF at a statistically significant level. Table 7 presents the DIF values subject to the gender of the items in Cluster05 of the mathematics subtest.

Table 7. DIF Values Subject to Gender of the Items in Cluster05

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q42	Female	-1.75	Male	-1.92	.17	.96
Q44		1.54		1.32	.22	.75
Q45		-1.12		-1.17	.05	.30
Q46		.18		.50	-.32	-1.53
Q47		2.26		2.09	.17	.44
Q48		-1.26		-1.26	.00	.00
Q49		-.83		-.96	.13	.77
Q50		3.37		3.13	.23	.39
Q51		-1.72		-1.55	-.17	-1.02
Q52		-.43		-.38	-.06	-.30

It was observed when Table 7 was examined that the items in Cluster05 do not exhibit DIF subject to gender. Figure 5 presents the DIF change graph subject to the gender of the items in Cluster05.

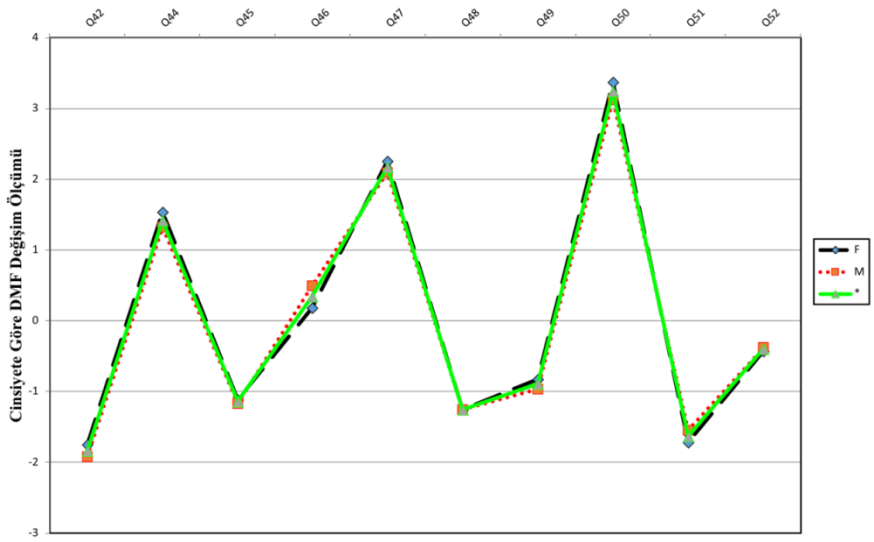

Figure 5. DIF Change Graph Subject to Gender of the 10 Items in Cluster05

It is understood when Figure 5 is examined that the items in Cluster05 do not display DIF at a statistically significant level. The DIF values subject to the gender of the items in Cluster06a of the mathematics subtest are presented in Table 8.

Table 8. DIF Values Subject to Gender of the Items in Cluster06a

Item	Focus Group	DIF Measurement	Reference Group	DIF Measurement	DIF Contrast	t Value
Q53		-1.74		-2.47	.73	3.82
Q54		.69		.27	.42	1.78
Q56		-.12		-.27	.16	.76
Q57		-1.60		-1.12	-.48	-2.55
Q58	Female	.48	Male	1.23	-.75	-2.94
Q59		2.85		2.82	.03	.06
Q60		-1.37		-1.45	.08	.42
Q62		.46		.59	-.13	-.55
Q63		.27		.61	-.35	-1.48

It is observed when Table 8 is examined that the item Q58 in Cluster06a operates in favor of females based on its DIF contrast (-.75) value, whereas item Q53 (.73) operates in favor of males. Figure 6 presents the DIF change graph subject to the gender of the items in Cluster06a.

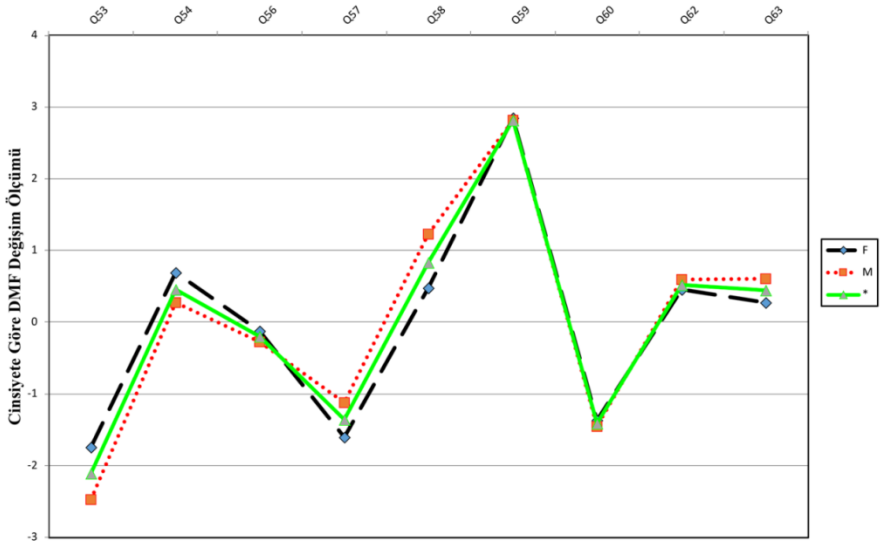

Figure 6. DIF Change Graph Subject to Gender of the Nine Items in Cluster06a

It can be observed when Figure 6 is examined that the female and male students stray away from the mean at the maximum level in items Q53 and Q58. In conclusion, it can be stated that items Q53 and Q58 exhibit DIF at a statistically significant level.

Statistical Regions Related DIF Findings of the Items in the Mathematics Subtest

This section focuses on the DIF values of the dichotomous (1-0) items in the PISA 2015 mathematics subtest. The DIF change graphs subject to statistical regions are presented for the items in the six clusters of the mathematics subtest. Figure 7 presents the change in DIF graph subject to the statistical regions for the nine items in Cluster01.

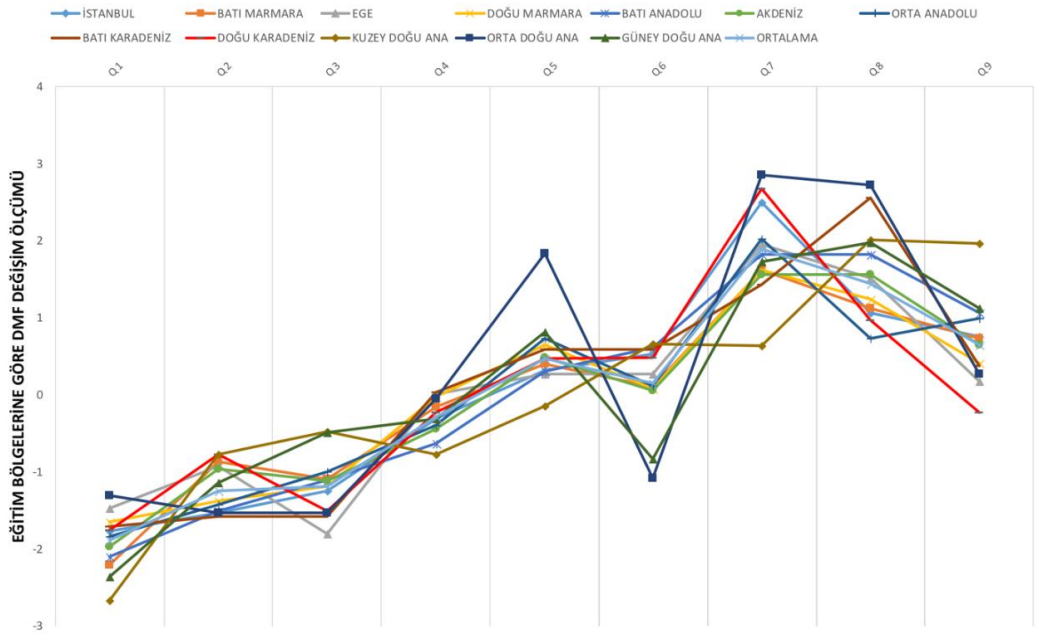

Figure 7. Change in DIF Subject to Statistical Regions for the Nine Items in Cluster01

In Figure 7, with the blue dotted line for İstanbul, Western Marmara with the orange checkered line for, Aegean with the grey triangle line, Eastern Marmara with the yellow crossed line, Western Anatolia with the blue starred line, the Mediterranean with the green dotted line, Central Anatolia with the navy blue perpendicular line, Western Black Sea with the brown line, Eastern Black Sea with the red line, Northeastern Anatolia with the brown checkered line, Central Eastern Anatolia with the navy blue squared line, Southeastern Anatolia with the green triangled line and the educational territory mean value with the blue crossed line. It is observed that items Q3, Q5, Q6, Q7, Q8 and Q9 exhibit the biggest change from among the items in Figure 1. It is seen from the DIF change graph of the items in Cluster01 subject to statistical regions that the maximum divergence from the mean value is in the Eastern Black Sea, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia region. Figure 8 presents the change in DIF of the eight items in Cluster02 subject to statistical regions.

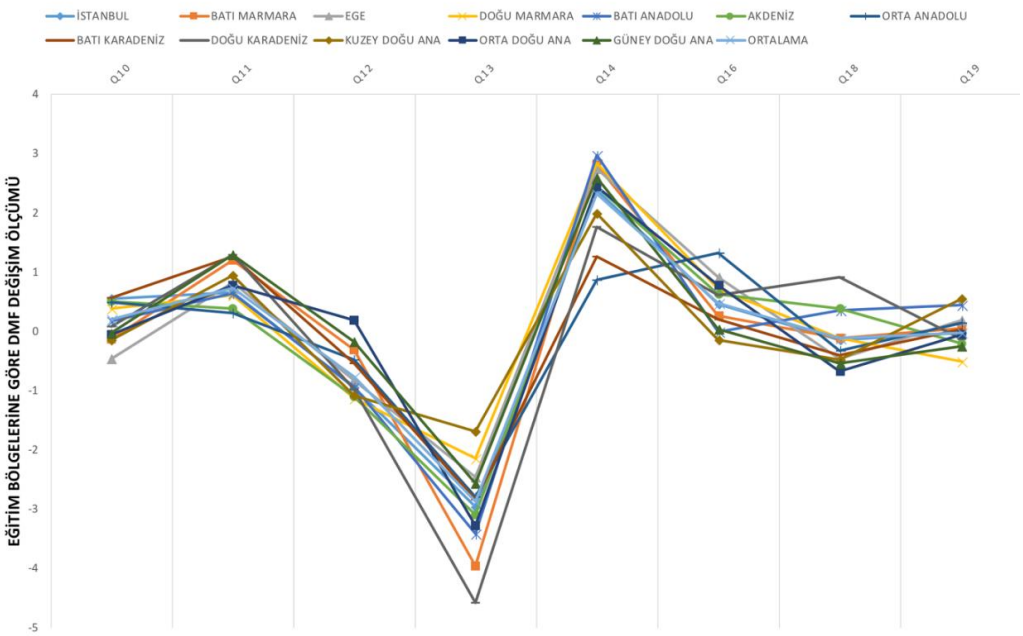

Figure 8. Change in DIF Subject to Statistical Regions for the Eight Items in Cluster02

Figure 8 illustrates that Q13, Q14, Q16 and Q18 items exhibit the highest rate of change. As can be seen from the change in DIF of the items in Cluster02 subject to statistical regions, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia are the regions that have diverged the most from the mean value. Figure 3 shows the DIF change graph of the 11 items in Cluster03 subject to statistical regions.

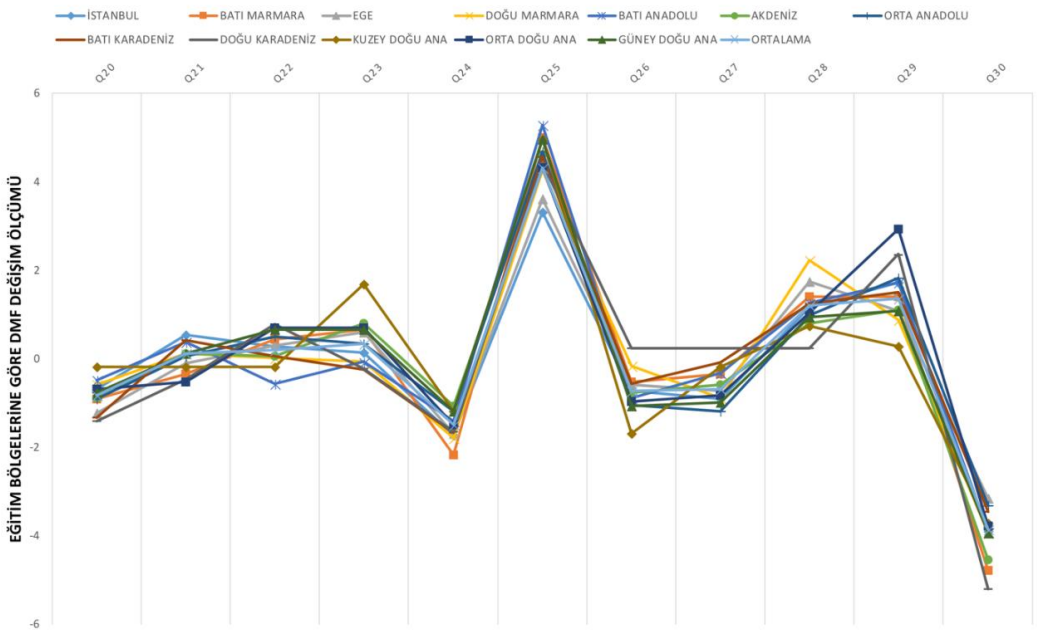

Figure 9. Change in DIF Subject to Statistical Regions for the Eleven Items in Cluster03

Figure 9 illustrates that Q23, Q25, Q26 and Q28 items exhibit the highest rate of change. As can be seen from the change in DIF of the items in Cluster03 subject to statistical regions, Eastern Marmara, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia are the regions that have diverged the most from the mean value. Figure 10 shows the DIF change graph of the eight items in Cluster04 subject to statistical regions.

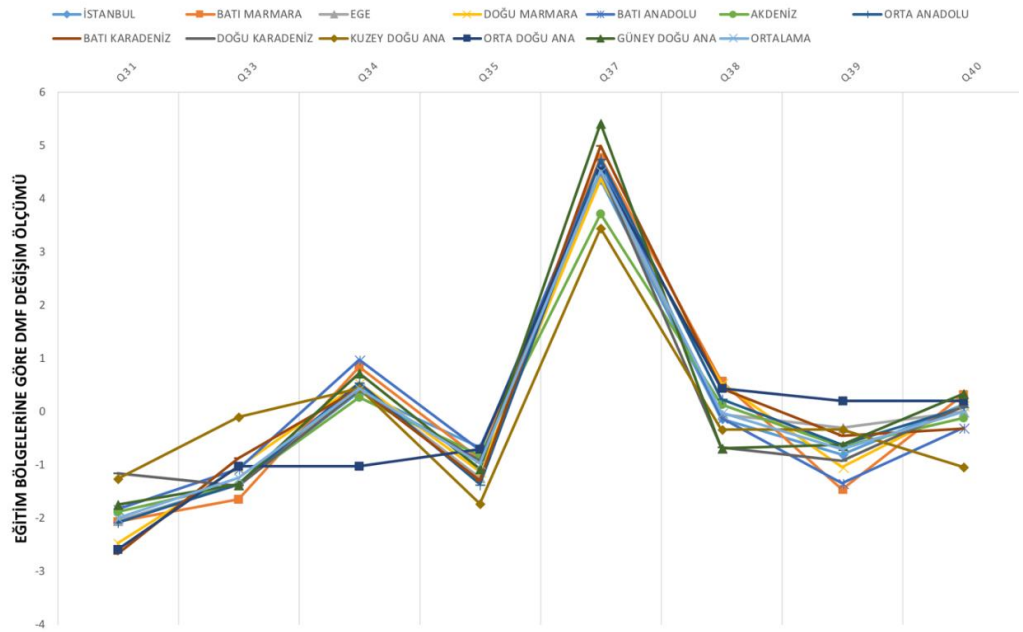

Figure 10. Change in DIF Subject to Statistical Regions for the Eight Items in Cluster04

It can be observed when Figure 10 is examined that Q33, Q34 and Q39 are the items from among the eight items of Cluster04, which display the highest rate of change subject to statistical regions. It is presented in the DIF change graph subject to regions for the items in Cluster04 that Northeastern

Anatolia and Central Eastern Anatolia are the regions that have diverged the most from the mean value. The results indicate that Northeastern Anatolia territory for items Q33 and Q40 and the Central Eastern Anatolia regions for items Q34 and Q39 have diverged from the mean value at a significant level. Figure 11 presents the DIF change subject to regions for the 10 items in Cluster05.

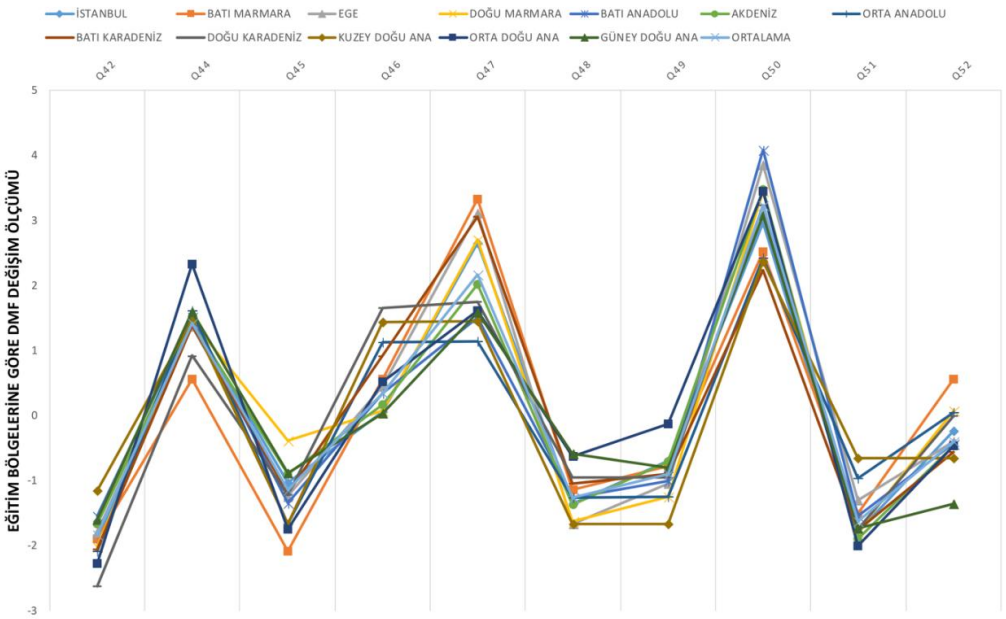

Figure 11. Change in DIF Subject to Statistical Regions for the Ten Items in Cluster05

Figure 11 reveals that the highest rate of change is observed in items Q44, Q46, Q47, Q49 and Q52. It can be seen from the graph showing the DIF change subject to statistical regions for the items in Cluster05 that the greatest divergence from the mean value has been observed in Western Marmara, Eastern Black sea, Northeastern Anatolia, Central Eastern Anatolia and Southeastern Anatolia regions. Figure 12 shows the DIF change graph of the nine items in Cluster06a subject to statistical regions.

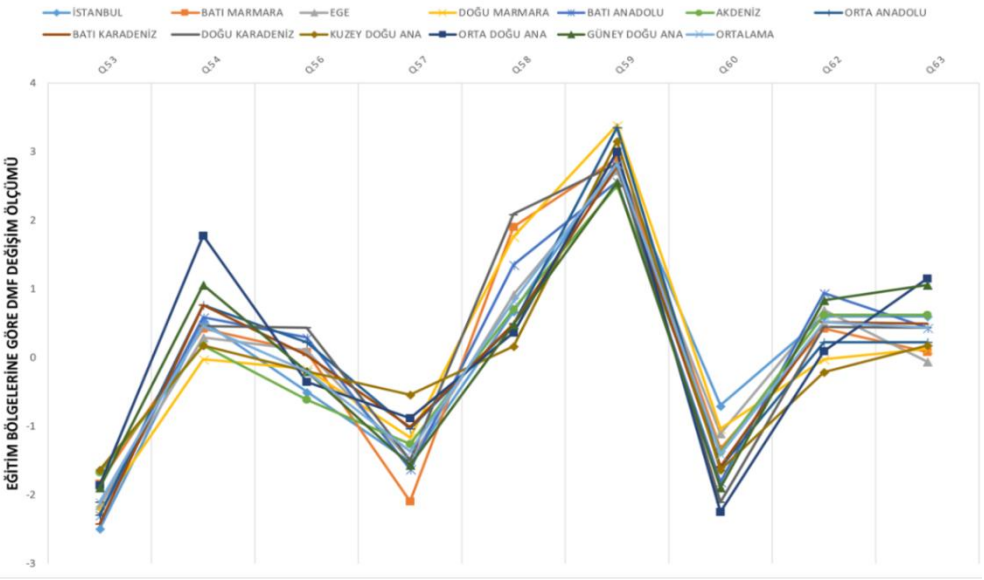

Figure 12. Change in DIF Subject to Statistical Regions for the Nine Items in Cluster06

It can be observed from Figure 12 that the highest rate of change in the graph is observed in items Q54, Q57, Q58, Q62 and Q63.

DISCUSSION and CONCLUSION

The aim of the present study was to determine whether the dichotomous (1-0) items in the PISA 2015 mathematics literacy subtest exhibit DIF subject to gender and statistical regions. IRT based Rasch model method was used for examining whether the items included in the study exhibit DIF or not.

Large scale examinations have a significant impact by way of their results on the education shareholders. It is important that the results of the examinations on which decisions related to individuals and the educational systems of countries are based contain minimum error. This will contribute to the accuracy of the decisions taken in accordance with the test results. Hence, it is expected that the tests applied in the field of education are free from bias. In other words, it is expected that the responses to the items in the examinations are not affected by factors such as gender, socioeconomic level, language, culture, territory, graduated school type, etc. excluding the abilities of the students. It is an important problem with regard to validity when the items operate for or against a certain group. Determining bias as a source of systematic error for examinations is also important for accountability.

Statistically significant DIF finding was observed subject to gender and statistical regions in the dichotomous (1-0) items of the mathematics subtest in PISA 2015 Turkey implementation. Statistically significant DIF findings were observed in five items in the mathematics subtest with regard to gender and in all items with regard to statistical regions.

Gender-based analyses of the items in the mathematics subtest of PISA 2015 revealed that item Q7 in Cluster01 operates in favor of males, whereas item Q8 operates in favor of females. Moreover, it was also understood that the item Q14 in Cluster02 operates in favor of males, whereas items Q58 and Q53 in Cluster06a operate in favor of females. It was understood as a result of the DIF analysis subject to gender for the 55 items of the mathematics subtest included in the study that five items exhibit. Of these items, two exhibited DIF in favor of females and three in favor of males. Demir and Köse (2014) carried out a study for examining whether the items included in PISA 2009 mathematics literacy subtest exhibit DIF subject to gender and culture. The study results put forth that two questions exhibit DIF subject to gender-based on the MH technique, three questions based on the LR technique and four questions based on the SIBTEST technique. In addition, DIF findings in favor of female students have been obtained as a result of the study by Akour et al. (2015) examining whether the PISA 2012 mathematics subtest results exhibit DIF or not. Atalay Kabasakal and Kıbrıslıoğlu Uysal (2017) conducted a study examining whether the PISA 2015 science subtest exhibits DIF subject to gender or not as a result of which it was observed that the number of items that exhibit DIF varies between two and six. Çıkrıkçı Demirtaşlı and Ulutaş (2015) carried out a study examining whether the items in the PISA 2006 science literacy subtest exhibit DIF subject to culture and gender or not. It was observed based on the DIF analysis subject to the item and item type that all multiple-choice items operate in favor of females, whereas two-thirds of the open-ended questions and a short response item operate in favor of males. It is observed especially in recent studies reporting gender DIF in large scale examinations that the number of items that exhibit DIF varies between three and six. The results obtained from these studies and the aforementioned literature findings are in accordance.

Statistically significant DIF findings were observed in all items when the results obtained from the analyses of the items in PISA 2015 mathematics subtest subject to statistical regions are examined. While at least 10 DIF cases were observed in the item based binary educational territory comparison, the maximum DIF cases observed were 38.

It is also very important for the implementation of the national and international tests in PISA 2015 mathematics subtest to take into consideration the impact of different demographic characteristics on the measurement results. It has been reported when the result indicating that the reasons for DIF in examinations carried out at the national level include variables such as gender and school type is taken

into consideration that it is inevitable for large scale examinations at the international scale such as PISA to include items that exhibit DIF (Bakan Kalaycıoğlu & Kelecioğlu, 2011). According to Sachse and Haag (2017), DIF can be observed due to the margin for error calculated for large scale examinations. In this regard, they have mentioned the need to reevaluate the methods used for calculating the standard error for national tendencies and taking into consideration the errors due to different points. Arikan, Van de Vijder and Yağmur (2018) expressed as a result of their study that less DIF is observed when tendency scores are used in DIF analyses.

It can be stated when the results of studies examining DIF subject to the gender of large scale examinations along with the results of the present study are taken into consideration that similar results have been attained. It is expressed when DIF subject to statistical regions is examined that different demographic characteristics should be taken into consideration and that different DIF prediction methods should be used. In addition, it is also observed that even though DIF description techniques yield similar results, they do not yield the same results due to the presence of algorithms and breakpoints at different classifications (Ardıç & Gelbal, 2017).

REFERENCES

Acar, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(1), 284-288.

Akalın, Ş. (2014). *Kamu Personeli Seçme Sınavı genel yetenek testinin madde yanlılığı açısından incelenmesi.* (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Akour, M., AL-Baddareen, G., H., A., & AL Duwairi, A. (2015). Exploring the jordanian gender gap in a large-scale assessment in mathematics. *Jordan Journal of Educational Sciences, 11*(1), 101-111.

Ardıç, E. Ö., & Gelbal, S. (2017). Cross-group equivalence of interest and motivation items in PISA 2012 Turkey Sample. *Eurasian Journal of Educational Research, 68*(2017), 221-238. doi: 10.14689/ejer.2017.68.12

Ariffin, S. R., Idrıs, R., & Ishak, N. M. (2010). Differential item functioning in Malaysian generic skills instrument (MyGSI). *Journal Pendidikan Malaysia, 35*(1), 1-10.

Arikan, S., Van de Vijver, F. R., & Yagmur, K. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education an Science, 4*(1), 69-82. doi: 10.21890/ijres.382936

Atalay Kabasakal, K., ve Kıbrıslıoğlu Uysal, N. (2017). Öğrenci özelliklerinin cinsiyete dayalı değişen madde fonksiyonuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(8), 373-390. doi: 10.21031/epod.333451

Ateşok Deveci, N. (2008). *Üniversitelerarası kurul yabancı dil sınavının madde yanlılığı bakımından incelenmesi.* (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bakan Kalaycıoğlu, D. (2008). *Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi*. (Doktora Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bakan Kalaycıoğlu, D., ve Kelecioğlu, H. (2011). Öğrenci Seçme Sınavının madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim, 36*(161), 3-13.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik Test Teorisi ve uygulanması.* Ankara: ÖSYM Yayınları.

Bekci, B. (2007). *Ortaöğretim Kurumları Öğrenci Seçme Sınavının değişen madde fonksiyonlarının cinsiyete ve okul türüne göre incelenmesi.* (Yüksek lisans Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Berberoğlu, G. (1995). Differential item functioning analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation, 21*(4), 439-456. doi: 10.1016/0191-491X(95)00025-P

Birjandi, P., & Amini, M. (2007). Differential item functioning (Test Bias) analysis paradigm across manifest and latent examinee groups (On the construct validity of IELTS). *Human Sciences*, 55 (Special Issue On Linguistics),153-172.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö., Karadeniz, Ş., ve Demirel, F. (2013). *Bilimsel araştırma yöntemleri.* (15. Baskı). Ankara: Pegem Akademi.

Camili, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* London: Sage Publications.

Çıkrıkçı Demirtaşlı, N., ve Uluştaş, S. (2015). A study on detecting of differential ıtem functioning of PISA 2006 science literacy ıtems in Turkish and American samples. *Eurasian Journal of Educational Research, 58,* 41-60 doi: 10.14689/ejer.2015.58.3

Demir, S., ve Köse, İ. A. (2014). Mantel-Haenszel, SIBTEST ve Logistik Regresyon yöntemleri ile değişen madde fonksiyonu analizi. *International Journal of Human Sciences, 11*(1), 700-714. doi: 10.14687/ijhs.v11i1.2798

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differantial item performance in mathematics achievement items. *Journal of Educational Measurement, 24*(2), 157- 166.

Ercikan, K., & Kim, K. (2009). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5*(1), 23-35. doi: 10.1207/s15327574ijt0501_3

Özer Özkan, Y., & Fincan, F. B. (2017). An Investigation of Item Bias in Free Boarding and Scholarship Examination in Turkey. *International Test Commission (ITC) 2016 Conference*. Canada

Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12(1), 29-51.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280-296.

Gök, B., Atalay Kabasakal K., ve Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 72-87. DOI: 10.21031/epod.64124

Hanna, G. (1986). Sex differences in the mathematics achievement of 8th graders in Ontario. *Journal for Research in Mathematics Education, 17*, 231-237.

Harris, A.M., & Carlton, S.T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Applied Measurement In Education*, 6(2),137-151.

Kıbrıslıoğlu Uysal, N., ve Atalay Kabasakal, K. (2017). Öğrenci özelliklerinin cinsiyete dayalı değişen madde fonksiyonuna etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(4), 373-390.

Kan, A., Sünbül, Ö., ve Ömür, S., (2013). 6.-8. Sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 9*(2), 207-222.

Karakaya, İ. (2012). Seviye belirleme sınavındaki fen ve teknoloji ile matematik alt testlerinin madde yanlılığı açısından incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(1), 215-229.

Karakaya, İ., ve Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi, *Eğitim ve Bilim, 37*(165), 348-362.

Kurnaz, F. B. (2006). *Peabody resim kelime testinin madde yanlılığı açısından incelenmesi*. (Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S. (2016). *Differential Performance on National Exams: Evaluating item and bundle functioning methods using English, Mathematics, and Science Assessments*. SAGE Open, 1-14.

Le, L. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing,* 9(2), 122-133. DOI: 10.1080/15305050902880769

Lyons-Thomas, J., Sandılands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education and Science Large- Scale Assessment Special Issue, 39*(172), 20-32.

MoNE, (2010). PISA 2009 Uluslararası Öğrenci Değerlendirme Programı Ulusal Ön Raporu, Ankara: MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MoNE, (2013). PISA 2012 Ulusal ön raporu, Millî Eğitim Bakanlığı, Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü, Ankara.

MoNE, (2015). PISA 2015 Uluslararası Öğrenci Değerlendirme Programı Ulusal Raporu, Ankara: MEB Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

OECD, (2015). OECD PISA Technical Report. OECD: http://www.oecd.org/pisa/data/2015-technical-report/ adresinden alındı. (Erişim Tarihi: 27.06.2018)

Osterlind, S. J. (1983). *Test item bias*. London: Sage Publications.

Öğretmen, T. (1995). *Differential item functioning (DIF) analysis of the verbal ability section of first stage of the university entrance examination in Turkey*. (Doctoral Dissertation, Middle East Tecnic University, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Öğretmen, T., ve Doğan, N. (2004). Öğretmen, T., & Doğan, N. (2004). OKÖSYS matematik alt testine ait maddelerin yanlılık analizi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 5* (8), 61-76.

Özer Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması*. (Doktora Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Özmen, D.T. (2014). PISA 2009 Okuma testi maddelerinin yanlılığı üzerine bir çalışma. *Eğitim Bilimleri ve Uygulama Dergisi*, *13*(26),147-165.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance", *Psychological Bulletin*, 114(3), 552-566.

Sachse, K. A., & Haag, N. (2017). Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Applied Measurement in Education*, *2*(30), 102-116. DOI: 10.1080/08957347.2017.1283315

Satıcı, K., ve Özer Özkan, Y. (2017). Temel eğitimden ortaöğretime geçiş sınavının (2014-Kasım) cinsiyet açısından madde yanlılığının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 254-274. DOI: 10.17860/mersinefd.305954

Suna, H.E. (2012). TIMSS 2007 Fen bilimleri testindeki maddelerin dil ve cinsiyet yanlılığı açısından incelenmesi. (Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Şenferah, S. (2015). *2010 Seviye belirleme sınavı matematik alt testi için değişen madde fonksiyonlarının ve madde yanlılığının incelenmesi*.(Doktora Tezi, Gazi Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed ıtem formats. *Applied Measurement in Education*, *25*(3), 246-280. DOI: 10.1080/08957347.2012.687650

Türkan, A. (2014). 2012- *Seviye Belirleme Sınavının Rasch modeline göre cinsiyet değişkeni açısından yanlılığının incelenmesi*. (Yüksek Lisans Tezi, Gaziantep Üniversitesi, Gaziantep). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Ulutaş, S. (2012). *PISA 2006 fen okuryazarlığı testindeki maddelerin yanlılık bakımından araştırılması*. (Doktora tezi, Ankara Üniversitesi, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Yalçın, S., ve Tavşancıl, E. (2015). TIMSS 2011 Fen uygulamasında cinsiyete göre farklılaşan madde fonksiyonunu açıklayan değişkenler. *Eğitim Bilimleri ve Uygulama*, *14*(27), 1-21.

Yurdugül, H., ve Aşkar, P. (2004). Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavının, öğrencilerin yerleşim yerlerine göre, diferansiyel madde fonksiyonu açısından incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *27*(2004), 268-275.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural Assessment. European *Journal of Applied Physiology*, *54*(2004), 119–135.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(12), 61-78. doi: 10.1207/s15326977ea0901&2_3

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP. *History Assessment Journal of Educational Measurement*, *26*(1), 55-66.

Appendix A. PISA 2015 Mathematics Literacy Form-Question and Cluster Matching Table