Identification of Differential Item Functioning on Mathematics Achievement According to the Interactions of Gender and Affective Characteristics By Rasch Tree Method

Münevver Başman [i]
Marmara University

Ömer Kutlu [ii]
Ankara University

Abstract

Mathematical knowledge and skills are needed to find solutions to the problems encountered in daily life. Although individuals are given the opportunity to receive equal education, it is seen that there are differences in the achievement of individuals. Individual-based factors can affect the achievement of individuals. One of the most important of these individual-based factors is the gender factor. It is important to examine the reasons behind the items of mathematics test showing the Differential Item Functioning (DIF) by gender. In this research, the interaction of gender and intrinsic motivation, instrumental motivation, self-efficacy, and anxiety variables on mathematics test items was examined in terms of DIF to understand the reasons of gender differences in the mathematical achievement of students who participated in PISA 2012. The study group of this research constituted 1084 students who participated in the application in Turkey, who answered booklets 3, 5 and 11 in the PISA 2012 mathematics literacy test. The data was analyzed by Iterative Hybrid Ordinal Logistic Regression (IHOLR) in the Lordif package program and Rasch Tree Method (RTM) in Psychotree package program and items showing DIF according to gender were determined. According to the findings, some mathematics test items showed DIF according to gender. It was found that items also showed DIF according to gender and intrinsic motivation interaction, and gender and self-efficacy interaction. It was observed that status of items showing DIF changed according to a certain threshold value of the girls' intrinsic motivation and self-efficacy score. It was found that mathematics items did not show DIF according to gender and instrumental motivation interaction, and gender and anxiety interaction. As a result, it was observed that status of items showing DIF according to gender could change according to gender and affective characteristics interaction.

Keywords: Differential Item Functioning, Mathematical Literacy, PISA, Rasch Tree Method

DOI: 10.29329/ijpe.2020.241.14

----

[i] **Münevver Başman,** Research Assist Dr., Educational Science, Marmara University, ORCID: 0000-0003-3572-7982

Correspondence: munevver.rock@gmail.com

[ii] **Ömer Kutlu,** Assist. Prof. Dr., Measurement and Evaluation, Ankara University

INTRODUCTION

Scientific and technological developments affect societies. Societies follow developments in the world and design their education accordingly. In today's rapidly developing world, the knowledge and skills expected from individuals can change. It is not enough for individuals only to have knowledge and they are expected to use the knowledge they have in daily life.

Societies want to nurture individuals who solve problems in their life and raise individuals to sustain their life. The fulfillment of these requirements which the societies expect from individuals requires them to have many basic knowledge and skills. One of these skills is the skill of mathematics. Mathematical knowledge and skills given to individuals increase the importance of mathematics teaching. Recent reforms in mathematics education are described with published documents in 1989, 1991, 1995 and 2000 by the National Council of Teachers of Mathematics (NCTM). It is stated in Assessment Standards for School Mathematics, one of the documents, that individuals should receive mathematics education equally regardless of gender, race, ethnicity and socioeconomic status (NCTM, 1995).

There are differences in the achievements of individuals who are considered to have mathematics education under equal conditions. Differences in mathematics achievement are also mentioned in international assessment projects. One of these projects is Programme for International Student Assessment (PISA). Since gender equality in education is important in terms of social justice and human rights, it has always been one of the most popular areas in international reports and studies. When examining gender differences in mathematics, boys show higher achievement than girls in 28 out of 31 participating countries in PISA 2000 and in 38 out of 40 countries in PISA 2003 (OECD, 2000 and 2004). In PISA 2000 and PISA 2012, in 38 out of 65 countries, boys show higher success than girls, while in only 5 countries this situation is reverse. In the remaining 22 countries, the achievements of girls and boys are similar (OECD, 2014). In PISA 2003 Turkey data, there are significant differences of 15 points in favor of boys in terms of mathematical literacy, whereas in PISA 2012 Turkey data, there are significant differences of 6 points in favor of boys.

When the differences according to gender are considered in terms of item type, it is seen that boys have higher achievement than girls in multiple-choice items (Bolger and Kellaghan, 1990; DeMars, 2000; Gipps and Murphy, 1994). Researches have shown that, as a cause of this condition, boys tend to take more risk and do not refrain from responding to items even if they are not sure, whereas girls prefer to leave the items blank (Ben-Shakhar and Sinai, 1991; Hanna, 1986). In addition, Liu and Wilson (2009a) illustrate that boys are superior to girls in complex multiple-choice mathematics items in PISA 2003. However, a different situation is observed in terms of constructed response items. DeMars (2000) and Gipps and Murphy (1994) found that girls show higher achievement than boys in constructed response items. As a reason for this situation, Bolger and Kellaghan (1990) and Bell and Hay (1987) state that girls express their thoughts more effectively because their language skills are higher. Lane, Wang and Magone (1996) emphasize that the mathematical processes required in constructed response tasks are explained in more detail by girls, whereas boys tend to focus on the results and tend to skip processes.

In PISA 2012, that the main subject of the project is the mathematics literacy, questionnaires consisting of some items on the attitudes and engagements with learning in mathematics have been applied. The questionnaire also aims to measure variables; *intrinsic motivation* (how much fun students have while learning mathematics), *instrumental motivation* (students' perceptions of using mathematics in their future studies and careers), self-efficacy (how much students trust their abilities in performing mathematical tasks), anxiety (how much students worry about their mathematics performance). It is thought that these variables have effects on mathematics achievement and affect individuals' differentiation in mathematics achievement (OECD, 2015).

Education Reform Initiative (ERI) (2014) published a report based on the PISA 2012 data for Turkey. In the report, they showed that girls were lower intrinsic motivation than boys, whereas girls were higher instrumental motivation than boys in mathematics. In terms of self-efficacy, it is stated that boys generally had higher self-efficacy than girls. Girls had higher self-efficacy than boys when answering the items included equations, whereas girls had lower self-efficacy than boys when answering the items included calculating the gasoline consumption rate of the car. They also stated that increasing self-efficacy beliefs of individuals with low self-efficacy may have an important role in preventing gender differences in mathematics achievement. When the role of anxiety variable in the gender differences in mathematics achievement is examined, they found that anxiety did not play an important role in the gender differences in mathematics achievement. However, socioeconomic status and type of program had an important role in mathematics success according to findings which were stated in their reports.

As can be seen, the achievements of individuals can be affected by psychological characteristics. The success of individuals in test items can be affected by secondary factors instead of the ability levels measured (Vi-Nhuan, 1999). There are three major sources of test bias for a particular group. The first is the bias that focuses on the content of the test. For example, items in the test may contain words which may be in favor of a group. The second is external bias. It refers to factors such as gender, race, language, the attitude of the individual, test anxiety, success, motivation and self-esteem. In addition, the type of items (such as multiple-choice, constructed response items), test time and individuals speed for answering the items are also sources of external bias. The third source of bias is the inappropriate use of tests (bias or injustice in choice) in selection and placement tests (Diamond, 1976; Green, 1981; Shinyoung, 1992; as cited in Eid, 2002).

In order to determine the bias, DIF analyzes are performed first. The methods of detecting differential item functioning are collected under the titles of Classical Test Theory (CTT) and Item Response Theory (IRT). Ellis and Raju (2003) state that Mantel-Haenszel (MH), Logistic Regression (LR) and delta plot (SIBTEST) methods are under CTT methods, whereas Lord's chi-square, Raju's area measures and likelihood ratio test methods are under IRT methods. Afterwards, different methods of detecting DIF have emerged. The Iterative Hybrid Ordinal Logistic Regression (IHOLR) and Rasch Tree (RT) methods are used in this study are some of the other methods. While IHOLR method is used because of combining logistic regression with the properties of IRT, RT method is used in terms of allowing multiple variables to be considered together to determine DIF in the items. Another advantage of RT method is that it does not need to specify the focus and reference groups as a prerequisite compared to most methods of determining DIF. RT method addresses the parameters of the item in all covariates when determining the groups and identifies the groups according to the covariate that gives the strongest instability (the inconsistency of the item parameters in the groups). For example, if it is desired to determine whether there is DIF in terms of gender and intrinsic motivation, it can be differentiated according to gender and then differentiated according to intrinsic motivation score. Whereas the covariate that gives the strongest item instability is gender, intrinsic motivation score is the second strongest one.

In addition, RT method has a superior feature in determining cut score than other methods. In methods that use predefined groups with continuous variables, arithmetic mean or median value is preferred as cutting points. In RT method, while grouping continuous variables, the value that gives the highest item parameter difference as the cutting point is considered. For example, some items show DIF by students' intrinsic motivation scores. When determining the focus and reference group, instead of the arithmetic mean or the median value of the intrinsic motivation scores of individuals, the cut-off point where the parameter difference is highest in RT method is taken into account. When the arithmetic mean or median value is chosen as the cut-off point, this selection is an arbitrary choice and may differ from the actual parameter difference which indicates the strongest parameter change. This may cause the actual parameter difference to be hidden by another cut-off point (Strobl, Kopf & Zeileis, 2015).

In this study, Rasch tree method is used because it is a new method in determining the individual traits behind DIF and it has some advantages from other methods. The aim of this study is to examine the mathematics items included in the PISA 2012 application within the context of the differential item functioning, depending on the interaction of motivation (intrinsic and instrumental), self-efficacy and anxiety variables with gender. For this purpose, the following questions were sought:

Do PISA 2012 mathematics items show the differential item functioning with respect to;

1. gender,

2. a combination of the covariates gender and intrinsic motivation,

3. a combination of the covariates gender and instrumental motivation,

4. a combination of the covariates gender and self-efficacy and

5. a combination of the covariates gender and anxiety.

METHOD

Research Model

The research is a descriptive research model in order to determine whether PISA 2012 mathematics test items show DIF according to the predefined variables and, it aims to describe the current situation as it exists (Karasar, 2010).

Study Group

PISA 2012 application includes 15-year-old students from 65 countries. 4848 students attend PISA 2012 Turkey application. These students are chosen randomly at 176 schools from 57 provinces representing the 12 regions determined by the Statistical Region Units Level 1 (Ministry of Education-MEB, 2015). The study group in this research consists of 1084 students participated in PISA 2012 mathematics applications from Turkey and answered only numbered 3, 5 and 11 test booklets from 13 test booklets. These booklets are preferred because the rate of missing data in these three booklets in Turkey application is less than other booklets.

Data Analysis

The study is carried out on 84 items in booklets 3, 5 and 11. The number of items in each booklet ranges from 11 to 37. IRT assumptions, whether speed test, unidimensionality, local independence and model-data compliance are examined. It is found that the test is not a speed test and it provides one-dimensional and local independence.

In order to determine whether PISA 2012 mathematics test items show gender-based DIF, the data are analyzed by RT method included in the psychotree package, and IHOLR method included in Lordif package in the R program. The data are analyzed with RT method in the Psychotree package program in the R program to determine DIF items according to combinations of the covariates gender and intrinsic motivation, self-efficacy, and anxiety variables.

The likelihood ratio chi-square test at significance level of .01 for IHOLR method, 5% differences in β_1 coefficient from Models 1 and 2 as a practically meaningful effect (Crane et al. 2004), and magnitude measures (ΔR^2) at least .035 are taken as the DIF determination criterion. Jodoin and Gierl (2001) indicate DIF levels as ΔR^2 <.035 DIF is absent or negligible, $.035 \leq \Delta R^2 < .070$ DIF is moderate, $\Delta R^2 \geq .070$ DIF is large. The significance level of .05 is considered as the DIF determination criterion for RT method.

FINDINGS

The findings obtained from the analysis of the data are given considering the order of the research questions.

Findings about whether PISA 2012 mathematics items show DIF by gender

In the DIF analysis performed by logistic regression likelihood ratio method, the 5th item (PM446Q01), 11th item (PM828Q01), 19th item (PM923Q01) and 25th item (PM995Q03) in the booklet 3, and the 26th item (PM955Q03) and 30th item (PM982Q04) in the 11th booklet show DIF by gender. In the booklet 5, there is no DIF by gender. Findings of DIF analysis are given in Table 1.

Table 1. DIF Analysis Findings Determined by IHOLR

Item number	Gender						
	Uniform DIF		Non-uniform DIF		Total DIF effect		Differences in regression coefficient
	p(χ^2_{12},1)	ΔR^2	p(χ^2_{23}, 1)	ΔR^2	p(χ^2_{13}, 1)	ΔR^2	%Δβ
5	.000	.0416	.0465	.0084	.000	.05	.0969
11	.000	.041	.1539	.0057	.000	.0468	.101
19	.000	.0254	.4602	.0011	.0016	.0266	.0061
25	.0042	.0166	.0296	.0096	.0016	.0262	.0063
26	.0033	.1245	.7641	.0013	.0127	.1258	.1105
30	.1663	.0039	.000	.0223	.0016	.0262	.0025

As can be seen in Table 1, DIF for 19th, 25th and 30th items can be negligible because of their effect sizes and differences in regression coefficient (ΔR^2 <.035, %Δβ <.05). For 5th and 11th items, DIF is moderate (.035 > ΔR^2 <.070, %Δβ > .05) and DIF for 26th item (ΔR^2 > 0.070, %Δβ > .05) is large. The properties of these items are presented in Table 2.

Table 2. Properties of DIF items determined by IHOLR

Item Number	Item Format	Content	Context	Process	Advantage
5	Constructed Response	Change and Relationships	Scientific	Formulate	Girls
11	Constructed Response	Change and Relationships	Scientific	Employ	Girls
19	Multiple Choice	Quantity	Scientific	Employ	Boys
25	Multiple Choice	Quantity	Scientific	Formulate	Boys
26	Constructed Response	Uncertainty and data	Societal	Employ	Boys
30	Multiple Choice	Uncertainty and data	Societal	Formulate	Boys

As shown in Table 2, the items which are in favor of girls are constructed response items, and the items which are in favor of boys (except for the 26th item) are multiple choice items. In constructed response items, girls express their ideas more effectively and their language skills are higher than that of boys. In multiple choice items, boys can take more risks while answering the items even they are not sure the correct answers.

The 25th item in the test is based on real life situations and requires using mathematical knowledge. The OECD (2015) report shows that girls are more successful than boys in solving mathematical problems similar to ordinary problems, but they are less successful than boys in defining a problem that can be encountered in everyday life as a mathematical problem in PISA (OECD, 2015). These findings can be shown as an important reason for the fact that item 25 shows DIF in favor of boys.

Although the 26th item is a constructed response item, it shows DIF in favor of boys. The fact that item 26 is in favor of boys is likely to be due to the higher achievement of boys in items that include probabilities, statistical events and situations.

The instability statistic values and p values for the booklet number 3 as a result of DIF analysis with RT are given in Table 3.

Table 3. Instability Statistical Value and p Value for Booklet Number 3 by Gender

Covariate	Instability	Node 1
Gender	Statistical value	.000
	p value	.000

As seen in Table 3, the gender as covariate is considered because the instability statistical value is significant (p<.05). This indicates that some items show DIF by gender. Item difficulty parameters of DIF-displaying items are given in Table 4 and the tree condition is shown in Figure 1.

Table 4. Difficulty Parameter Values of Items by Gender

Gender	Items						
	5.	10.	11.	16.	19.	21.	25.
Boy	-1.190	-3.676	1.598	-3.922	-0.579	1.830	-0.700
Girl	-2.101	-3.083	0.729	-4.755	0.346	2.833	0.035

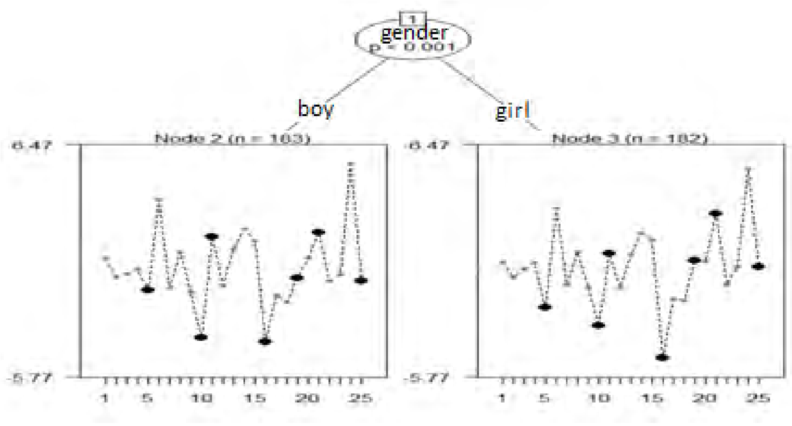

Figure 1. Rasch Tree for Booklet Number 3 by Gender

As seen in Figure 1, 5th, 10th (PM800Q01), 11th, 16th (PM918Q01), 19th, 21st (PM923Q04) and 25th items show DIF between girls and boys. Strobl, Kopf and Zeileis (2015) indicate that high value of item means item is difficult whereas low value of item means item is easy. The 5th, 11th and 16th items are in favor of girls, while the 10th, 19th, 21st and 25th items are in favor of boys. The properties of these items are presented in Table 5.

Table 5. Properties of DIF items determined by RT

Item Number	Item Format	Content	Context	Process	Advantage
5	Constructed Response	Change and Relationships	Scientific	Formulate	Girls
10	Multiple Choice	Quantity	Personal	Employ	Boys
11	Constructed Response	Change and Relationships	Scientific	Employ	Girls
16	Multiple Choice	Uncertainty and data	Societal	Interpret	Girls
19	Multiple Choice	Quantity	Scientific	Employ	Boys
21	Constructed Response	Change and Relationships	Scientific	Formulate	Boys
25	Multiple Choice	Quantity	Scientific	Formulate	Boys

As shown in Table 5, 5th and 11th items are constructed response items and there are DIFs in favor of girls. The 19th and 25th items are multiple choice items and DIFs are found in favor of boys. In addition, according to gender by RT, it is found that 10th, 16th and 21st items show DIF. When the 10th item is examined, it is seen that the item is in multiple choice format and in favor of boys.

The 16th item is a multiple choice item and shows DIF in favor of girls. When the 16th item is examined, it is seen that the matter is related to music groups in terms of context. 21st item is a constructed response item and shows DIF in favor of boys. When the 21st item is examined, it is seen that the item is related to the consumption of vehicles in terms of context and it is an item which is identified with male roles. The comments about 16th and 21st are given in more detail below, in terms of intrinsic motivation and self-efficacy.

Instability statistic values and p values for the booklet 5 and 11 as a result of DIF analysis by using RT are given in Table 6.

Table 6. Instability Statistical Value and p Value for Booklet Number 5 and 11 by Gender

	Booklet 5		Booklet 11	
Covariate	Instability	Node 1	Instability	Node 1
Gender	Statistical value	54.1061216	Statistics	47.73566018
	p value	0.3037928	p value	0.05921611

When Table 6 is examined, there is no difference between boys and girls (p>.05). This shows that Booklet 5 and 11 do not contain DIF items by gender.

Findings about whether PISA 2012 mathematics items show DIF by a combination of the covariates gender and intrinsic motivation

According to the gender and intrinsic motivation interaction using RT for booklet 3, instability statistics values and p values are given in Table 7 as a result of DIF analysis.

Table 7. Instability Statistical Value and p Value for Booklet Number 3 by Gender and intrinsic motivation

Covariates	Instability	Node 1	Node 3
Gender	Statistical value	.000	
	p value	.000	
Intrinsic Motivation	Statistical value	51.95183561	50.60542014
	p value	0.04980374	0.03482394

As seen in Table 7, gender is found to have the strongest instability value and it is observed that there is a significant instability (p <.05). The intrinsic motivation covariate is considered after the covariate of gender and it is found to have significant instability (p<.05). This indicates that items that exhibit DIF by gender also show DIF according to the intrinsic motivation covariate. Item difficulty parameters of DIF items are given in Table 8 and the tree is given in Figure 2.

Table 8. Difficulty Parameter Values of Items with respect to a combination of the covariates gender and intrinsic motivation

Gender	Intrinsic Motivation Score	Item 5	Item 10	Item 11	Item 16	Item 19	Item 21	Item 25
Boy		-1.190	-3.676	1.598	-3.922	-0.579	1.830	-0.701
Girl	≤10	-1.931	-2.822	0.448	-4.637	0.649	3.006	-0.227
	>10	-2.253	-3.452	1.438	-4.836	0.283	2.656	-5.863

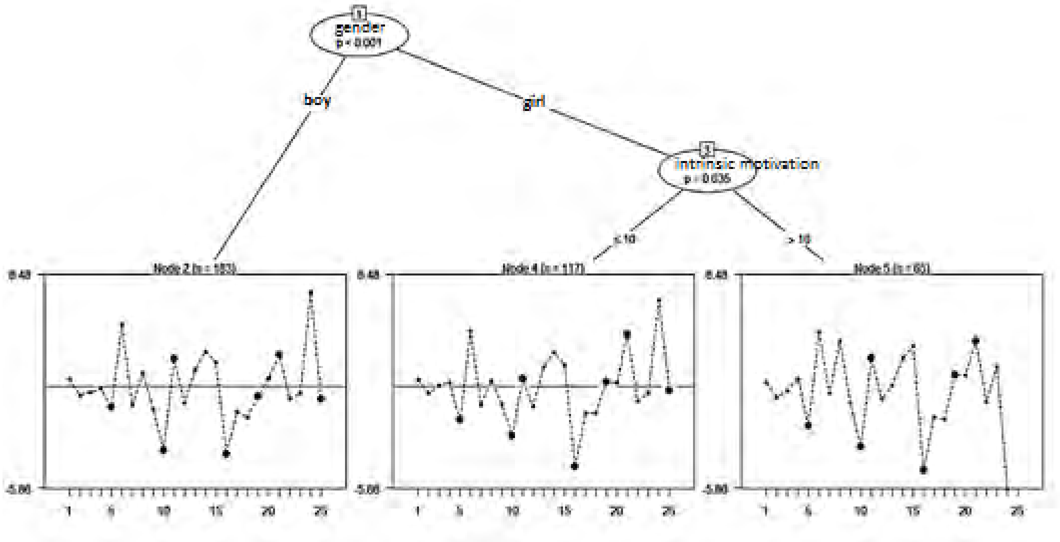

Figure 2. Rasch Tree for Booklet Number 3 by Gender and Intrinsic Motivation

When Figure 2 is examined, it is seen that there are some differences between boys and girls who have intrinsic motivation scores of 10 points and less than 10 points and girls who have that of more than 10 points. In this case; (i) item 5 which is in favor of girls is easier for them with an intrinsic motivation score of more than 10 points, (ii) item 10 which is in favor of boys is more difficult for girls with intrinsic motivation score of 10 points and less than 10 points, (iii) item 11 which is in favor of girls is easier for girls with an intrinsic motivation score of 10 points and less, (iv) item 16 which is in favor of girls is easier for girls with an intrinsic motivation score of more than 10 points, (v) item 19 which is in favor of boys is more difficult for girls with intrinsic motivation score of 10 points and less, (vi) item 21 which is in favor of boys is more difficult for girls with intrinsic motivation score of 10 points and less, (vii) item 25 which is in favor of boys is more difficult for girls with intrinsic motivation score of 10 points and less. It is also observed that item 25 is in favor of girls with an intrinsic motivation score of more than 10 points compared to boys.

While 5th and 11th constructed response items are in favor of girls; 10th, 19th and 25th multiple choice items are in favor of boys. This situation may be explained by that boys' intrinsic motivations are higher than girls' on multiple choice items. While boys may exhibit more risky and responsive behavior even they are not sure the answers because of their intrinsic motivation, girls tend to leave blank instead of responding to the items.

In terms of context, item 16 is related to music groups. Simpkins, Fredricks, Eccles and Simpkins-Chaput (2012), in their longitudinal studies, have modeled on how families' beliefs affect the performance of adolescent children. It is observed that families support their boys in sports activities, computer use, mathematics and science, and support their girls in music. In this case, they state that the girls give more importance to music and they are interested in music more. The 16th item may show DIF in favor of girls because of the higher interest of high school girls than boys and their intrinsic motivation.

The 21st item, the constructed response one, shows DIF in favor of boys. In the Education Reform Initiative-ERI (2014) report, it is stated that the self-efficacy of boys is higher than that of girls in items identified with boys' roles. Item 19 and item 21 relate to gasoline consumption of vehicles. Considering the finding in the ERI report, the sample status used in the relevant items may have increased the self-efficacy perception of boys. An increase in self-efficacy perceptions may increase intrinsic motivation so that items may show DIF in favor of boys.

Findings about whether PISA 2012 mathematics items show DIF by a combination of the covariates gender and instrumental motivation

According to the gender and instrumental motivation interaction using RT for booklet 3, instability statistics values and p values are given in Table 9 as a result of DIF analysis.

Table 9. Instability Statistical Value and p Value for Booklet Number 3 by Gender and Instrumental Motivation

Covariates	Instability	Node 1
Gender	Statistical value	.000
	p value	.000
Instrumental Motivation	Statistical value	45.9628447
	p value	0.1933588

As seen in Table 9, gender is found to have the strongest instability value and it is observed that there is a significant instability (p<.05). The instrumental motivation covariate is considered after the covariate of gender and it is found that it does not have significant instability (p>.05). This may be interpreted as there is no instrumental motivation variable among the possible sources of items that show DIF according to gender.

Findings about whether PISA 2012 mathematics items show DIF by a combination of the covariates gender and self-efficacy

For the booklet used in PISA 2012, according to the interaction of gender and self-efficacy perceptions using RT, instability statistic values and p values are given in Table 10 as a result of DIF analysis.

According to the gender and self-efficacy interaction using RT for booklet 3, instability statistics values and p values are given in Table 10 as a result of DIF analysis.

Table 10. Instability Statistical Value and p Value for Booklet Number 3 by Gender and Self-Efficacy

Covariates	Instability	Node 1	Node 3
Gender	Statistical value	.000	
	p value	.000	
Self-Efficacy	Statistical value	38.2705845	50.6945309
	p value	0.6463309	0.0340753

As seen in Table 10, gender is found to be the covariate, giving the strongest instability value and it is a significant instability (p <.05). The self-efficacy covariate is considered after the covariate of gender and it is found to have significant instability. This indicates that items that display DIF according to gender also show DIF according to self-efficacy variable. Item difficulty parameters of DIF items are given in Table 11 and the tree is shown in Figure 3.

Table 11. Difficulty Parameter Values of Items with respect to a combination of the covariates Gender and Self-Efficacy

Gender	Self-Efficacy Score	Item 5	Item 10	Item 11	Item 16	Item 19	Item 21	Item 25
Boy		-1.190	-3.676	1.598	-3.923	-0.579	1.830	-0.701
Girl	≤13	-1.930	-2.482	0.332	6.474	0.536	-0.792	-0.035
	>13	-2.482	-3.206	1.014	-4.463	0.281	3.566	-5.773

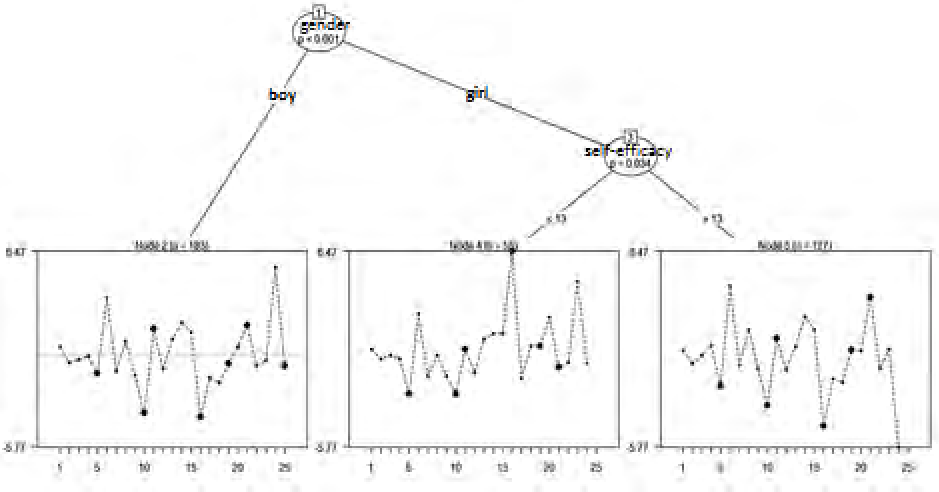

Figure 3. Rasch Tree for Booklet Number 3 by Gender and Self-Efficacy

Figure 3 shows that there are differences among boys and girls who have self-efficacy scores of 13 points and less than 10 points and girls who have that of more than 13 points. In this case; (i) 5th item which is in favor of girls is easier for girls with a self-efficacy score of more than 13 points, (ii) item 10 which is in favor of boys is more difficult for girls with a self-efficacy score of 13 points and less than 13 points, (iii) item 11 which is in favor of girls is easier for girls with a self-efficacy score of 13 points and less than 13 points, (iv) item 16 which is in favor of girls is easier for girls with a self-efficacy score of more than 13 points. In addition, it was observed that 16th item is in favor of boys due for girls with self-efficacy score of 13 points and less than 13 points. (v) item 19 which is in favor of boys is more difficult for girls with self-efficacy score of 13 points and less than 13 points (vi) item 21 which is in favor of boys is more difficult for girls with self-efficacy score of more than 13 points. In addition, it was observed that the 21st item was in favor of girls with self-efficacy score of 13 points and less. (vii) Item 25 which is in favor of boys is more difficult for girls with self-efficacy score of 13 points and less. In addition, it is seen that the 25th item is in favor of girls with a self-efficacy score of more than 13 points.

While the 5th and 11th items, the constructed response ones, are in favor of girls, 10th, 19th and 25th items, the multiple choice ones, are in favor of boys. The DIF source in multiple choice items can be explained by that the intrinsic motivation and the self-efficacy perception of boys is higher than that of the girls. Individuals with high self-efficacy perceptions enjoy working with mathematical tasks, and exhibit more persistent behaviors to perform the task (Zimmerman, 2000).

The 21st item, constructed response one, shows DIF in favor of boys. In the ERI (2014) report, it is stated that the expression of calculating the gasoline consumption rate of a car is explained by the boys' roles and the self-efficacy of the boys in these items is higher than the girls. Items 19 and 21 may be interpreted as increasing the self-efficacy perception of boys as they are related to gasoline consumption and thus may have shown DIF in favor of boys.

Findings about whether PISA 2012 mathematics items show DIF by a combination of the covariates gender and anxiety

According to the gender and anxiety interaction using RT for booklet 3, instability statistics values and p values are given in Table 12 as a result of DIF analysis.

Table 12. Instability Statistical Value and p Value for Booklet Number 3 by Gender and Anxiety

Covariates	Instability	Node 1
Gender	Statistical value	.000
	p value	.000
Anxiety	Statistical value	42.1323713
	p value	0.3864828

As seen in Table 12, gender is found to have the strongest instability value and it is observed that there is a significant instability (p<.05). The anxiety covariate is considered after the covariate of gender and it is found that it does not have significant instability (p>.05). This may be interpreted as there is no anxiety variable among the possible sources of items that show DIF according to gender.

CONCLUSION AND DISCUSSION

In this study, firstly, it is examined whether the mathematics items in booklet number 3, number 5 and 11 in PISA 2012 application show DIF with respect to gender. Item 5, 11, 19, and 25 show DIF according to gender analyzing by both IHOLR and RT, while item 10, 16 and 21 show DIF according to gender only analyzing by RT. Item 5 and 11 are in favor of girls in both IHOLR and RT, while items 19 and 25 are in favor of boys in both IHOLR and RT. In general, constructed response items show DIF according to gender by both IHOLR and RT are in favor of girls while multiple choice items show DIF according to gender by both IHOLR and RT are in favor of boys. However, some multiple choice items are found to show DIF in favor of girls and some constructed response items are favorable to boys. It can be concluded that the reasons underlying the display DMF by gender are not solely dependent on item properties like items' contents, contexts and thinking processes.

Liu and Wilson (2009b) have reached the conclusion that boys is slightly superior to the girls and multiple choice items are in favor of boys in PISA 2000 and PISA 2003 mathematics literacy. Bolger and Kellaghan (1990) examine the gender differences in school mathematics achievement in their studies, and state that boys are more successful in multiple choice items while girls are more successful in constructed response items. Garner and Engelhard (2009) also have found that multiple choice items show DIF in favor of boys, constructed response items show DIF in favor of girls. When they consider the items as mathematical content, they have found that algebra-containing items show DIF in favor of girls, and geometry and measurement, probability and statistics, data analysis and proportion-containing items show DIF in favor of boys. In this study, it is seen that in general, constructed response items are in favor of girls and multiple choice items are in favor of boys. In addition, boys are more successful in terms of uncertainty and data-containing items. This is consistent with the results of Bolger and Kellaghan (1990) and Garner and Engelhard (2009).

In this study, it is seen that some multiple choice items are in favor of boys and some constructed response items are in favor of girls. This suggests that it is not enough to explain the sources that underlie the items showing DIF by gender only with the properties of items, and that the affective characteristics of individuals may be DIF sources by gender. In this context, items showing DIF by the gender also show DIF a combination of the covariates gender and intrinsic motivation. The threshold value of the girls' intrinsic motivation score is found to be 10 points. This shows that there are differences in success among girls with an intrinsic motivation score of 10 points and less and that of more than 10 points. Similarly, it is seen that DIF items by gender also show DIF by the interaction of gender and self-efficacy perceptions. The threshold value of the girls' self-efficacy score is found to be 13 points. This shows that there are differences in success among girls with a self-efficacy score of 13 points and less and that of more than 13 points. On the other hand, there is no DIF according to a combination of gender and instrumental motivation and a combination of gender and anxiety. In the ERI (2014) report, it is stated that the expressions associated with the boys' roles in the items increase the self-efficacy of boys and that items can be in favor of boys. It is also stated in the report that it is important to encourage girls to be motivated by mathematics and to increase their self-confidence. In this study, it has been stated that some items in favor of boys have increased the self-efficacy perception of boys because they contain statements that are identified with boys' roles. In addition, it can be seen that the success of girls can be increased by increasing their intrinsic motivation and self-efficacy perceptions.

As a result of this study, examining gender-DIF sources of mathematics items not only in terms of item properties, but also in terms of affective properties, it can contribute to writing more qualified items. As another result of this study, it can be seen that affective characteristics may cause differences in mathematics achievement. Teachers can be reminded that both boys and girls are supported by affective characteristics based on their ability to succeed in mathematics. It can be suggested that teachers should carry out their lessons in this context to support experiences to meet the needs of boys and girls students and support these courses with appropriate materials.

In this study, mathematics intrinsic motivation, mathematics instrumental motivation, mathematics self-efficacy and mathematics anxiety which are some affective characteristics are examined as gender-DIF sources of mathematics items. Apart from these, it can be suggested to examine the affective features such as mathematical self-perception, mathematical behavior, and problem solving determination, mathematics working ethics, and openness to problem solving. In addition, this study focuses on the underlying causes of items that display DIF by gender in mathematics. The variables such as socioeconomic status and school type can affect mathematics achievement. Therefore, it may be suggested that researchers investigate the underlying causes of DIF items such as socioeconomic status and school type.

REFERENCES

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology, 57*, 212-220.

Ben-shakar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 77-92.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165-174.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1)*,* 55-77.

Eid, G. K. (2002). *Gender, ethnicity, and language influences on differential item functioning in the SAT*. Unpublished Ph.D. thesis, Ohio University, United States.

Ellis, B. B., & Raju, N. S. (2003). Test and item bias: What they are, what they aren't, and how to detect them. Educational Resources information center (ERIC).

ERI. (2014). *Türkiye PISA2012 analizi: Matematikte öğrenci motivasyonu, özyeterlik, kaygı ve başarısızlık algısı.* Eğitim Reformu Girişimi, Araştırma Notu, Sabancı Üniversitesi.

Garner, M., & Engelhard, G. (2009). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied measurement in education, 12*(1), 29-51.

Gipps, C., & Murphy, P. (1994). *A fair test: Assessment, achievement and equity*. Buckingham: Open University Press.

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

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Karasar, N. (2010). *Bilimsel araştırma yöntemleri.* 21. Baskı. Ankara: Nobel Yayın Dağıtım.

Kopf, J. (2013). *Model-based recursive partitioning meets item response theory: New statistical methods for the detection of differential item functioning and appropriate anchor selection.* Unpublished doctoral dissertation, LMU München: Faculty of Mathematics, Computer Science and Statistics.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27, 31.

Liu, O. L., & Wilson, M. (2009a) Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing, 9*(1), 20-40.

Liu, O. L., & Wilson, M. (2009b). Gender differences in large scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*, 164-184.

MEB. (2015). *PISA 2012 araştırması ulusal nihai rapor*. Milli Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. Ankara: İŞKUR Matbaacılık.

National Council of Teachers of Mathematics. (1995). *Assessment Standards for School Mathematics*. Reston, VA: NCTM.

OECD. (2000). *Knowledge and Skills for Life: First Results from PISA 2000*. OECD Publishing.

OECD. (2004). *Learning for Tomorrow's World: First results from PISA 2003*. OECD Publishing.

OECD. (2014) *PISA 2012 results: what students know and can do—student performance in mathematics, reading and science,* vol. I, OECD Publishing.

OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence,* OECD Publishing.

OECD. (2016). Data base PISA-2012. Retrieved from Web: https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm

Simpkins, S. D., Fredricks, J., Eccles, J. S., & Simpkins-Chaput, S. (2012). Charting the Eccles' Expectancy-Value Model from parents' beliefs in childhood to youths' activities in adolescence. *Developmental Psychology*, *48*, 1019-1032.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316.

Vi-Nhuan, L. (1999). *Identifying Differential item functioning on the NELS:88 History Achievement Tests.* CSE Technical Report 511. Stanford University, CA: National Center for Research & Evaluation.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*, 82-91.