International Education Studies; Vol. 12, No. 8; 2019
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Detecting Gender Differences in PISA 2012 Mathematics Test with Differential Item Functioning

Ozen Yildirim[1]

[1] Faculty of Education, Pamukkale University, Denizli, Turkey

Correspondence: Ozen Yildirim, Faculty of Education, Measurement and Assessment Department, Pamukkale University, Denizli, Turkey.

Received: April 3, 2019 Accepted: May 16, 2019 Online Published: July 29, 2019

doi:10.5539/ies.v12n8p59 URL: https://doi.org/10.5539/ies.v12n8p59

Abstract

The measurement tool not measuring the specific construct has a validity problem. Individuals based on the results obtained from this type of tool should not be evaluated. The purpose of this study was to examine the differentiated item functioning and item bias of mathematics items in the Programme for International Student Achievement 2012 assessment for gender using two-level hierarchical generalized linear model, logistic regression and experts' opinions. Also differentiated item functioning sources (anxiety, interest and self-efficacy) at student level were tested. The current study was created under take into account of quantitative and qualitative methods. It was conducted with 1458 students selected from 166 schools of Turkey sample. The results reveal that hierarchical generalized linear models approach is more conservative than logistic regression approach. When the student level variables were added to the model as potential sources, differentiated item functioning did not disappear for the three items. Also half of the experts argued that the items identified as in favor of boys are biased. Statements in the items and the context were given as the reasons for this bias.

Keywords: differential item functioning, hierarchical generalized linear model, item bias, logistic regression, validity

1. Introduction

Program for International Student Achievement (PISA) is a large scale assessment study conducted worldwide by the Organization for Economic Cooperation and Development (OECD). The purpose of the assessment is to determine students' daily life performances in mathematics, science and reading literacy. PISA results and reports are amongst the sources of reference that countries examine in order to organize their education policies. Although not PISA's aim, based on these results countries compare each other in terms of student performance and make critics.

Such large-scale assessments, which are conducted by being translated into different languages, should be reliable and valid. Otherwise, the interpretations about students made will be scientifically incorrect or incomplete. Wrong decisions can be made on students whose tests scores are not objective. Today, one of the issues discussed amongst test developers and psychometrics is the comparability of results from tests. For example, if two students with the same achievement level do not perform the same in a mathematics question, what may be the reason for this? Is there a problem with the validity of the test or is the performance affected by different variables?

PISA technical reports (OECD, 2012, 2014b) are provided on the validity and reliability of measurement instruments. However, the analyses explained in the technical report are limited. An in-depth examination of the tests measuring cognitive and affective properties in PISA, which provides a large data source, will be beneficial. In the present study, the construct validity of the mathematical cognitive domain test was examined for Turkey. For this, differential item functioning (DIF) of the items was examined according to gender, and the sources of DIF was examined.

Validity refers to what construct the test measures and how accurately the test measures that construct (Turgut & Baykul, 2014). The interpretations made based on findings obtained from a measurement tool whose validity was not proven will not carry weight. Although different methods are used to determine validity, differentiation in the probability of individuals with the same ability or achievement level in different groups (focus, reference)

responding to an item is examined using differential item functioning methods (Hambleton & Rogers, 1995). For example, the probability of giving the correct answer to the item is supposed to be equal for students with the same ability in the same group or in the different groups. If the students who are within the same population but are the member of separate groups, they can have a different probability of giving the correct answer. This is associated with item bias (Dogan, Hambleton, Yurtcu, & Yavuz, 2018; Zumbo, 1999). Examining item bias which characteristics is supposed to be discussed is important because item bias based on many different group characteristics (gender, language, socio-economic status, race, etc.) may occur (Millsap & Everson, 1993).

There are many analysis techniques based on traditional approaches. Many of them accept the assumption that the function of differentiating factors for individuals with the same characteristics is found in the same pattern. However, individuals are often clustered within different organizations such as class, school, country, and the behaviors of differentiating factors also vary in each organization. Therefore, using multilevel methods that do not ignore the relation between clusters is recommended for DIF determination (Adams, Wilso, & Wu, 1997; Kamata, 2001). In this study, two approaches were used. One of these approaches was Generalized Hierarchical Linear Models (HGLM) as multilevel model (Kamata, 1998, 2001). The other approach was logistic regression (LR) as traditional approach (Zumbo, 1999). Thus the results of the two approaches will be compared.

Expert opinions are vital to reveal the items' bias. The sensitivity review is to put forth the source of bias in the items. Taking into consideration the individuals with the same performance, experts examine the possibility of individuals responding to the items solely because the item contents are close to a specific group's experiences. In the study, the opinions of the experts who specialized in writing the test items on the related subject were taken.

How the probability of students responding to an item might change based on different variables can be determine by using HGLM. Also the disappearance of DIF in items can be examined. In the study, the variables of self-efficacy, mathematics anxiety and mathematics interest, which all believed to be effective on student achievement, were examined how changed magnitude of DIF.

1.1 Detecting DIF with HGLM

Data in the social sciences mostly has a clustered structure. Repeated measures are found in clusters made up of individuals whereas individuals are found in different organizational units. For example, the items in the test are answered by students and these students are in a class, and the school is made up of these classes. While the smallest unit of organization in this example is the item, the highest unit is the school. For this reason, the elements of each level in a hierarchy are commonly affected by the specific characteristics and experiences of the upper levels. In measurements the assumption of observations is independent of each other is not unacceptable. This will lead to loss of information and Type I and Type II error rates (Hox, 2002; Snijders & Bosker, 1999). Researchers suggest that hierarchical models that do not ignore this assumption are used in data analyses (Hox, 2002; Raudenbush & Bryk, 2002).

The DIF detection method used in the study is the proposed HGLM model for binary coded items by Kamata (1998, 2001). The reason for using HGLM is because HGLM takes into consideration the clustered data structures in large-scale assessments; gives information about the psychometric properties of the test; examines the sources of DIF by adding different characteristics to the model and because there is no need to separate the sample into two groups as reference and focus (Qian, 2011). HGLM model developed for binary coded items and allows missing data in responses during parameter estimation (Kamata, 1998, 2001). The first level of the HGLM model is the item level while the second level is the student level. Levels can be increased according to the purpose of the study. This study was designed based on the two-level HGLM approach. The log odd of "j" student's (in class "m") probability of correctly answering the item "i" is calculated without adding the group variable to the models. Then, Group variable (gender) in which the DIF will be examined is added to the second level, and the significance of this variable on the items is tested. In the third step, different student characteristics are added to this level, and possible DIF sources can be examined by looking at the change of magnitude of DIF.

1.2 Detecting DIF with Logistic Regression

Logistic regression approach is the one of the most recommended methods for determining DIF (Clauser & Mazor, 1998; Swaminathan &Rogers, 1990). DIF is divided into uniform and non-uniform. Uniform DIF occurs when no interaction between ability and group membership can be found. Non-uniform DIF occurs when interaction between ability and group membership can be found (Camilli & Shepard, 1994; Swaminathan & Rogers, 1990). Results of simulation-based studies revealed that LR approach was powerful at a comparable level in detecting uniform DIF than Mantel-Haenszel (MH) and Simultaneous Item Bias (SIB) methods and LR approach was quite powerful in detecting non-uniform DIF (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990).

In the LR process, item responses are used as dependent variables, and group variable (reference=1 and focus=0), total score calculated for each individual, group and total score-group interaction are included in the models as independent variables. With this method, the DIF based on the relationship between item response and total score is determined, whereas the type of DIF is determined with the testing of the group and group total score interaction (Zumbo, 1999). By comparing the regression coefficients and the fit indices of the three different models obtained by the addition of this independent variable to the model, information about item bias and the significance level of this bias is obtained.

1.3 Gender Differentiation in Mathematic Achievement

Numerous studies at national and international level have reported gender differentiation in mathematics performance. Analyses of NAEP data revealed a small but persistent gender gap in scores (Lubienski, McGraw, & Strutchens, 2004; McGraw, Lubienski, & Strutchens, 2006). Analyses of the Early Childhood Longitudinal Study showed no gender gap at the beginning of kindergarten. However, a gender gap began to form during the early elementary years. In fact, this gender gap favored boys by third grade (Fryer & Levitt, 2010; Robinson & Lubienski, 2011). Kane and Mertz (2012) examined the gender differences in countries' mathematical performances from the large-scale tests such as TIMSS and PISA, and found that these gender differences occurred in many countries.

Turkey's PISA results indicate obvious achievement differences between boys and girls. Girls show a better performance than boys in science and reading whereas boys perform better in mathematics than girls (OECD, 2014a). While the study emphasized that there was no significant difference between the genders in terms of achievement in elementary school, they also revealed that the gender difference in mathematics and science began in middle school. In middle school, girls' mathematics achievement, in particular, displayed an apparent drop (Sanders & Nelson, 2004). In many studies conducted in recent years, examination of the reasons behind the differences in achievement has gained importance. It was emphasized that different characteristics such as race, socio-economic status, affective and cognitive characteristics could be the reason for this difference (Harris & Herrington, 2006; Zimmerman & Kitsantas, 2005). For example, Gallagher et al. (2000) reported that boys were more flexible in applying problem-solving strategies than girls. Girls abided by the in-class learning steps more than boys. They showed this as the source of differentiation between the performances.

In the education reforms and specifically in reforms in the mathematics education, the importance of raising individuals who can think, understand and establish cause and effect relationships and who have mathematical ability is emphasized for the future (National Research Council, 2001) To be able to raise individuals with these qualities, the teaching and the materials used in teaching are expected to be valid, reliable and appropriate to the student level. It is important to determine the factors leading to differences amongst individuals with similar performances. In this way, students can be evaluated fairly, and the right decisions can be made about them.

1.4 Student Characteristics as Potential Sources

Mathematical anxiety, one of the affective characteristics discussed in this study, involves anxiety and tension felt by the student in his or her daily and academic life while dealing with subjects based on mathematics (Vahedi & Farrokhi, 2011). Having difficulty in mathematics can be caused not only by the student's inadequacy in mathematical learning but also by mathematics anxiety (Maloney & Beilock, 2012; Vukovic, Kieffer, Bailey, & Harari, 2013). Mathematics anxiety can have a negative effect on mathematics performance. The student with anxiety has less confidence in solving math problems and feels inadequate. This negatively affects the student's future mathematics performance (Vahedi & Farrokhi, 2011). The results of studies on Western societies revealed that girls have math anxiety at higher levels compared to boys (Else-Ques, Hyde, & Linn, 2010; Frenzel, Pekrun, & Goetz, 2007; Goetz et al., 2013). Goetz et al. (2013) determined that girls had lower perceived competence than boys even though they had the same average grades in mathematics. Parallel to this finding, girls reported higher levels of anxiety than boys.

The student's interest towards a subject involves indicators such as willingly participating in the class and believing that the subject is important for his or her career. As a student's interest level towards a subject increase, his or her ability to use cognitive processes also increases. According to the results of Lubinski and Benbow (2006)'s longitudinal study on how to increase students' interest towards mathematics, when the students' interest increases, the importance they gave to mathematics also increases. These students believe that mathematics has an important place in their future careers. While studies revealed very few gender gaps in mathematics interests and mathematics achievement between boys and girls in elementary school (Wigfield et al., 1997), studies showed prominent gender gap when children reach adolescence and girls began to report less interest towards mathematics. In their study, Frenzel et al. (2007) found that boys were more interested in mathematics than girls.

According to the results of studies conducted by using international data like TIMSS and PISA, boys had more positive attitudes towards mathematics than girls in almost all the participating countries (Else-Quest et al., 2010; Liu &Wilson 2009). Girls consider mathematics as a male-dominated field, and they prefer the fields with more verbal cognitive processes.

Students' mathematics self-efficacy, as another affective characteristic, is also discussed in this study. Self-efficacy involves a person's task-specific, instead of general, perception, belief and expectation about one's own ability to achieve in a specific area (Bandura, 1997). Self-efficacy is generally considered as a predictive characteristic that helps the student with his or her choices, effort in a subject and commitment in academic issues (Bandura, 1977). Individuals with higher levels of self-efficacy attempt more cognitively challenging problems are more effortful, use productive problem-solving strategies and persist longer (Pajares, 1996; Pajares & Graham, 1999). Studies reported that boys have higher self-efficacy in mathematics than girls whereas girls show higher self-efficacy in language arts (Meece, Glienke, & Burg, 2006; Siegle & Reis, 1998).

The purpose of this study was to examine the differential item functioning of items in the PISA 2012 mathematics test for the gender using HGLM and LR approaches and to obtain expert opinions on item bias. Furthermore, potential sources were tested by HGLM.

2. Method

2.1 Research Design

The research consists of two parts: quantitative and qualitative. In the quantitative part, it was study to determine DIF items by using HGLM and logistic regression techniques. In addition, the potential sources of items with DIF were determined. In the qualitative dimension of the study, expert opinions were evaluated qualitatively and descriptively evaluated in order to determine item bias.

2.2 Sample

This study was conducted with data from the PISA 2012 Turkey sample. The study sample consisted of 1458 students selected from 166 schools. In PISA, students do not respond to all the mathematics items in the test. Instead, they respond only to the specific items in the booklet they are given. During the analysis process, each student's response distributions of the items to be examined should be available. Also, the researcher wanted to reach as many as common items from different booklets. The Turkey sample consisted of 4848 students. However, the study was conducted with the 1458 students who had the same common items in their booklets. 51% of the sample was male (748), while 49% (710) was female students.

In the second part of the study, the item bias was evaluated by 14 experts. The opinions of the experts who received training on developing items were taken. The characteristics of the experts are presented below.

Table 1. The characteristics of the experts

Experts	Gender	Occupation	Area	Professional year
E1	Female	Teacher	Mathematics	3
E2	Female	Academician	Mathematics	18
E3	Male	Academician	Mathematics	24
E4	Female	Teacher	Mathematics	11
E5	Female	Teacher	Mathematics	36
E6	Female	Teacher	Mathematics	18
E7	Female	Academician	Measurement and evaluation	14
E8	Male	Academician	Measurement and evaluation	18
E9	Female	Teacher	Mathematics	6
E10	Female	Academician	Mathematics	9
E11	Male	Academician	Mathematics	20
E12	Female	Academician	Measurement and evaluation	7
E13	Female	Academician	Measurement and evaluation	10
E14	Male	Academician	Measurement and evaluation	8

The evaluation of items was conducted by four men and 10 women. Five of them were mathematics teachers, four of them were mathematics education specialists and five of them were measurement and evaluation specialists. The expert with the least occupational experience had three years of experience, and the expert with the most

occupational experience had 36 years of experience.

2.3 Data and Collection

In the study, the data from PISA 2012 mathematics cognitive domain achievement test items and student questionnaires from Turkey were used. The data were taken from the official website of the OECD, which carries out the PISA applications. The following briefly gives information based on the measurement tools and the properties it measures.

2.3.1 Math Test Items

In the study, the following criterion was taken into consideration during the selection of the items that would be examined:

- ✓ The hierarchical data structure should be appropriate,
- ✓ The sample size should be appropriate for data analysis,
- ✓ Items should be published.

According to these criteria, 13 mathematics items that were answered by 1458 students and that were included in booklet 1, booklet 3, booklet 4 and booklet 6, were analyzed. In terms of item content, four of the items were on space and shape, three on uncertainty and data, three on change and relations and three on quantity. All items were multiple choice items. In terms of cognitive level of items, five of them were on formulate level; six on employ level and two on interpret level. PISA 2012 item codes respectively: PM00FQ01 (item1), PM903Q01 (item2), PM903Q03 (item3), PM918Q01 (item4), PM918Q02 (item5), PM918Q05 (item6), PM923Q01 (item7), PM923Q03 (item8), PM923Q04 (item9), PM924Q02 (item10), PM995Q01 (item11), PM995Q02 (item12), PM995Q03 (item13). Since PISA questions are based on long or short texts, only one example is shown in the study (See Appendix A). All of these items are available at https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

2.3.2 Student Questionnaire

In current study, gender, math anxiety, math interest and math self-efficacy were obtained by the student questionnaire (OECD, 2014). The gender variable (boy and girl) on which DIF is being examined. The independent variables as possible sources of DIF were mathematics anxiety*(ANXMAT)*, mathematics interest(INTMAT) and mathematics self-efficacy (MATHEFF).

Mathematics anxiety (ANXMAT) was measured with five items. The items have four response categories: “strongly agree”, “agree”, “disagree” and “strongly disagree”. Higher difficulty corresponds to higher level of anxiety. The scale’s Cronbach Alfa is 0.82 for Turkey.

Mathematics interest (INTMAT) was measured with four items.). The items have four response categories: “strongly agree”, “agree”, “disagree” and “strongly disagree”. Higher difficulty corresponds to higher level of interest. The scale’s Cronbach Alfa is 0.89 for Turkey.

Mathematics self-efficacy was measured with eight items. The items have four response categories: “very confident”, “confident”, “not very confident” and “not at all confident”. Higher difficulty corresponds to higher level of confidence. The scale’s Cronbach Alfa is 0.82 for Turkey.

2.3.3 Expert Opinion Questionnaire

During the development of the questionnaire studies with expert opinions on item bias were reviewed (Demirtaşlı & Ulutaş, 2015; Yalçın, 2015) and the questionnaire was developed under cover of these scales. The experts were given a questionnaire with 13 items. Their opinions were collected for content validity. Corrections were made according to four expert opinions.

The 13 items in the questionnaire asked whether each test item has bias or not and whether the bias is according to gender if there is a bias. Furthermore, 14 experts were asked to examine item bias in terms of item content, context, question format and visuals.

2.4 Data Analysis

In the study, HGLM and logistic regression was used. Evaluations were made using the common findings. Uniform and non-uniform DIF were detected.

For logistic regression analysis, three-step regression equation was created based on total score and group variable, and group variable and total score interaction according to student performance. Then, items with DIF were detected by testing the differences between the chi-square values of models ($\Delta\chi^{2\,(3\text{-}1)}$ and $\Delta\chi 2^{\,(2\text{-}1)}$) and the

regression coefficients in each step.

The data obtained from the tests performed on students in the field of education shows a multi-level structure. Students in the same classroom and school can be affected by similar characteristics. Since this assumption is not taken into consideration in analyses based on traditional regression method, there is a probability of higher risk error. The PISA sample is selected based on a stratified sample structure. Therefore, this study used Hierarchical generalized linear model (HGLM) method. According to this approach, the item is at the lower level of the hierarchical level. The probability of student answering the items in the test is affected from each other. Also, each answer is influenced by the student's characteristics. The likelihood of responding to item may vary according to the student characteristics.

In HGLM, only items that are coded as dummy were added to Level 1. In Level 2, two-category group indicator variable was added to each equation. The significance of the gender variable in the model was examined.

In the third step, Item difficulty indexes and students' ability levels of DIF detected items were determined using Zero Rasch model (intercept-only) in order to determine the characteristics of DIF detected items. The content, context and cognitive characteristics of items were examined.

Finally, student's mathematics self-efficacy, mathematics anxiety and mathematics interest, which are all second level variables, were added to model. How magnitude of DIF changed when these variables are added in the model was discussed.

Expert opinions were collected qualitatively through questionnaires. Expert evaluation of the bias of the items and the bias of the same items according to gender were determined by frequency analysis. The experts were also evaluated the items terms of content, context, question format and visuals. The comments of them are quoted directly and were given samples. During the analysis of qualitative data, the researcher carried out the whole study himself, since no scoring, comparison of opinions or a content analysis based on themes were conducted. During evaluation of the data, the steps taken in the researches where expert opinions based on item bias were taken were followed (Çepni, 2011; Demirtaşlı & Ulutaş, 2015; Kalaycıoğlu & Kelecioğlu, 2011; Yalçın, 2015).

3. Results

In DIF determination, logistic regression approach was first used. Table 2 includes the results of LR. It was found that six items have DIF. Uniform and non-uniform DIF were identified by examining the differences between the chi-square values $\Delta\chi^{2\,(3\text{-}1)}$ and $\Delta\chi2^{\,(2\text{-}1)}$ in Table 2. Accordingly, six items (item 1, item 2, item 3, item 4, item 7 and item 9) displayed uniform DIF. Four of them are significant at the level of 0.01. It was also determined that the same six items presented non-uniform DIF. Four items are significant at the level of 0.01. The significance levels of the items with DIF were examined based on differentiation between R^2s. Gierl, Jodoin, & Ackerman (2000) expressed 0.035 as the criterion value. The values were smaller than 0.035. Therefore, the magnitude of DIF was not significant.

Table 2. Results of DIF with LR

Item	$\Delta\chi^{2\,(3\text{-}1)}$	$\Delta\chi2^{\,(2\text{-}1)}$	$\Delta R^{2\,(3\text{-}2)}$	$\Delta R^{2\,(2\text{-}1)}$
Item1	11.806**	8.682**	0.003	0.006
Item2	14.035**	5.623*	0.005	0.004
Item3	6.493*	5.428*	0.002	0.004
Item4	8.324*	6.922**	0.002	0.010
Item5	0.213	0.09	0.000	0.000
Item6	0.009	0.000	0.000	0.000
Item7	35.111**	32.491**	0.002	0.022
Item8	4.787	3.673	0.001	0.003
Item9	13.103**	10.79**	0.004	0.014
Item10	0.821	0.763	0.000	0.000
Item11	0.226	0.135	0.000	0.000
Item12	2.616	2.123	0.002	0.009
Item13	0.773	0.003	0.000	0.000

**p<0.01; *p<0.05.

DIF findings obtained using the HGLM method is presented in Table 3. In HGLM analysis, one item needs to be defined as the reference item. In the study, item 13 was identified as the reference item. According to Table 3, two items were significant at the 0.01 level, and one item was significant at the 0.05 level. In other words, these items displayed DIF, while item 7 and item 9 works in favor of boys, item 4 works in favor of girls.

Table 3. Results of DIF with HGLM

Item	β Coefficient	S.E
Intercept2	-0.129	0.138
Item1	0.224	0.185
Item2	0.109	0.182
Item3	0.168	0.199
Item4	0.506*	0.242
Item5	0.066	0.173
Item6	0.044	0.175
Item7	-0.530**	0.174
Item8	-0.208	0.178
Item9	-0.826**	0.264
Item10	0.082	0.169
Item11	0.028	0.174
Item12	0.461	0.500

**p<0.01; *p<0.05.

The common items 4, 7 and 9 were determined to display DIF based on logistic regression and HGLM.

When the structural and content characteristics of the items were examined: The items' difficulty values were 0.715, 0.806, 1.855, -2.885, -0.795, -0.955, 0.602, 1.013, 2.845, 0.004, 0.436, 4.815 and 0.616 respectively. Item 5 has an easy item difficulty degree and Item 7 has a medium item difficulty degree whereas Item 9 has a difficult item difficulty degree.

In favor of girls, Item 4 is an item requiring reading graphics where the student needs to compare different situations. Cognitive characteristic of the item is interpretation. The item's subject is uncertainty. The item's context is social.

In favor of boys, Item 7 aims to measure student's ability to calculate percentages which is based on real life situations. Cognitive characteristic of the item is employment. The item's mathematical subject is determining quantity. The item's context is scientific.

In favor of boys, Item 9 aims to measure student's problem-solving ability on subjects of reduction of fuel consumption and cost which are based on real life situations. Cognitive characteristic of the item is formulation. The item's mathematical subject is change and relations. The item's context is scientific.

In the last step of the study, a second model was developed in order to determine the sources of DIF and how magnitude of DIF changed according to selected variables. At this time, the gender variable, which was in the first model, was kept constant and the other three variables were added to the model (ANXMAT, INTMAT, MATHEFF). Table 4 shows how the beta coefficient changed when other variables were added to the model.

Table 4. Changing magnitude of DIF by potential sources

Item	DIF Coefficient	S.E
Intercept2	-0.083	0.136
Item1	0.211	0.184
Item2	0.161	0.185
Item3	0.237	0.203
Item4	0.485*	0.242
Item5	0.059	0.174
Item6	0.054	0.175
Item7	-0.516**	0.174
Item8	-0.215	0.177
Item9	-0.806**	0.267
Item10	0.119	0.171
Item11	0.052	0.177
Item12	0.442	0.505

**p<0.01; *p<0.05.

The findings revealed that the Items 4, 7 and 9 which were found to display DIF according to gender still displayed DIF when the three variables were added to the model. When student's anxiety level increased, the probability of the student answering correctly decreased for Item 2 (-0.242, se=0.113 p=0.033), Item 3 (-0.253, se=0.124, p=0.042) and Item 11 (-0.444, se=0.111, p=0.000). Student's interest only increased in Item 7's correct answering probability (0.260, se=0.113, p=0.022). In PISA, the student's self-efficacy for mathematics is determined by the student's confident level. As the student's MATHEFF level increased in Item 1 (-0.308, se=0.122, p=0.012) and Item 8 (-0.321, se=0.115, p=0.006), the probability of the student answering the question correctly decreased whereas in item 11 (0.260, se=0.126, p=0.038), this probability increased.

After the comparison of the methods used in determining DIF, 14 experts were asked about item bias according to gender. The expert opinions are given in Table 5.

Table 5. HGLM, LR and expert opinions' distribution in terms of item bias for gender

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13
Statistical results													
HGLM				X(G)			X(B)		X(B)				
LR	X	X	X	X			X		X				
Expert Opinions													
E1	X(B)						X(B)						
E2													
E3	X(B)	X(G)	X(G)	X(G)			X(B)		X(B)				
E4				X(G)	X(G)	X(B)	X(B)	X(B)	X(B)	X(G)			
E5													
E6													
E7		X(G)	X(G)				X(B)		X(B)				
E8		X(G)	X(G)				X(B)	X(B)	X(B)	X(G)			
E9													
E10							X(B)		X(B)				
E11													
E12	X(B)			X(G)		X(G)				X(G)			
E13	X(B)	X(G)	X(G)				X(B)	X(B)	X(B)	X(G)			
E14	X(B)								X(B)				

Note. E: Expert, G: Girls, B: Boys.

Five experts detected item bias in favor of boys for item1. When the experts were asked about the cause of bias, they stated that the visual of a house plan provided in the question and the subject of the question were more

relevant to the boys. Examples of their direct quotations are as follows:

"I think that male students may be more advantageous because they are more interested in design, engineering and today's computer games and because the question is about their interest area." (E1)

"The real estate business and the visual of a house plan given in the question are more familiar for boys." (E13)

Four experts determined item bias for item2 and stated that the bias was in favor of girls. According to experts, the profession the question was about appealed more to the girls, and this made the question work in their favor. Examples of their direct quotations are as follows:

"The pictures in the question. The nursing profession is seen appropriate for women in the culture they are in." (E7)

"Nursing is accepted as a profession performed by women in our society. The recruitment of male nurses has been happening recently." (E8)

Item 3 was found to be biased in favor of girls by four experts. This question is a continuation question prepared using the visual in Item 2. Similar to Item 2, experts believed that the question created a bias in favor of girls because the nursing profession is considered a more appropriate profession for women in general. Some parts of their direct quotations are as follows:

"Because the question is based on the related visual, the problem is reflected as the problem of a female nurse. So, I think girls will be more active in solving this question." (E3)

"Because the context of the question is based on nursing." (E1)

Item 4 was found to be biased in favor of girls by three experts. The experts believed the item was biased because it had a visual graph. Another reason for the bias was the question's content. Some parts of their direct quotations are as follows:

"I think girls are better at visual questions like graph interpretation." (E4)

"The questions are about music. The girls may be more familiar with the content because 15-year old girls are more interested in music, and they follow the latest albums." (E12)

Item 7 was found to be 50% biased in favor of boys. Experts emphasized that the concepts found in the question are more familiar to boys. Some parts of their direct quotations are as follows:

"Male students' interest in vehicles and speed." (E1)

"It would be more interesting for both genders if the expressions used in the question like tanker and cargo ship are interesting for the girls." (E7)

"The concepts in the question like tanker and diesel fuel are more meaningful and familiar for boys. There may even be girls who don't even know what diesel fuel is." (E13)

Item 9 was found to be 50% biased by the experts. The item bias was in favor of boys. The question was about navigation. Experts considered navigation as a profession more familiar to boys. Furthermore, they emphasized that the concepts given in the question are more relevant to boys' daily lives. Some parts of their direct quotations are as follows:

"Because concepts like speed and vehicle attract the attention of male students more." (E4)

"Since navigation is a profession that can be associated with men, it is believed that it may cause bias in favor of boys." (E8)

"The concepts of money, trade and fuel given in the question are the subjects more familiar to boys." (E13)

4. Discussion and Conclusions

The present study aims to examine the construct validity of the Turkish PISA 2012 mathematics test. For this purpose, whether the probability of answering the 13 items in the test was differentiated or not was tested for gender using hierarchical generalized linear models and logistic regression approach. According to the literature, the DIF sources were examined in terms of the anxiety, interest and self-efficacy variables. The disappearance of DIF was also examined. Then, expert opinions were taken to prove the item bias.

The study results have been reached after taking certain steps. First, logistic regression approach and then HGLM was used. Six items displayed DIF in logistic regression approach whereas three items presented DIF in HGLM. DIF in logistic regression approach, a traditional approach, is based on the interpretation of chi-square values. Chi-square value is affected by the sample size (Zumbo, 2009). Therefore, more items might have been

found biased in logistic regression approach. The reason why the HGLM method was preferred is because the sample structure in large scale test applications such as PISA shows a stratified structure. When the probability of answering the item is taken into consideration, the probability of a student answering an item might be affected by the student's own characteristics (Raudenbush & Bryke, 2002). Here, while the first level is made up of answers given to the items, the second level is made up of student characteristics. Another important point in the analysis based on DIF is to examine the DIF size. When DIF sizes of the six items were examined according to the related criteria, the magnitude of DIF was found small. However, three items were found to display DIF in both of the analysis methods. For this reason, the analysis was continued and the steps were followed on these three common substances.

One of the three items with DIF was in favor of girls. The content of this item was based on reading and interpreting graphs about the change in music bands' album sales throughout the year. The vast majority of experts did not determine the item as being biased. Those who found the item biased believed that there might be a problem with the context of the item.

The items in favor of boys were mostly based on daily life. They were also at problem-solving application and formulation level. Gallagher et al. (2000) found that girls in particular strictly follow learning-based steps in the classroom, and they insist on applying these steps when they are faced with a problem situation. In addition, male students are more flexible in using problem-solving strategies and can try different ways to solve a problem. Studies conducted by Geary (1996) and Halpern (2000) support the previous finding. According to these studies, girls are successful in algebra, which is a part of mathematics that has a certain structural language. However, their performances tend to decrease in large-scale tests which include high risk and in which the content differ from the education program. Half of the experts determined these two items to be biased rather than the item in favor of girls and emphasized the context of the items were more associated with male professions. While the items in favor of boys were at medium and hard difficulty level, they required problem-solving ability based on daily life. Studies have shown that male students are more advantageous in problem solving in geometry, cause-effect determination and spatial situations in nature (Geary, 1996; Hyde, Fennema & Lamon, 1990). In addition, as the item difficulty increases, they are more disadvantaged in terms of mathematics performance (Bielinski & Davidson, 2001; Penner, 2003).

As a result of the addition of mathematics anxiety, mathematics interest and mathematics self-efficacy variables to the model, DIF was not eliminated according to gender, instead only minor changes were observed.

Since the results of large-scale testing applications such as PISA, TIMSS (Trends in International Mathematics and Science Study) and PIRLS (The Progress in International Reading Literacy Study) lead countries' education policies, the tests used in these applications should be valid and reliable. Researchers can examine the psychometric properties of these large-scale tests by taking into account the different characteristics of the student or the country. Also DIF sources according to gender can be examined by variables at the teacher and school levels.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47-76. https://doi.org/10.2307/1165238

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*(1), 51-77. https://doi.org/10.1111/j.1745-3984.2001.tb01116.x

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Beverly Hills, CA: Sage.

Çepni, Z. (2011). *Değişen madde fonksiyonlarının SIBTEST, mantel haenszel, lojistik regresyon ve madde tepki kuramı yöntemleriyle incelenmesi* (Unpublished doktoral dissertation). Hacettepe University, Ankara, Turkey.

Clauser, B. E., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Demirtaşlı, N., & Ulutaş, S. (2015). A study on detecting differential item functioning of PISA 2006 science

literacy items in Turkish and American samples. *Eurasian Journal of Educational Research*, *58*, 41-60. https://doi.org/10.14689/ejer.2015.58.3

Dogan, N., Hambleton, R. K., Yurtcu M., & Yavuz, S. (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Science*, *13*(2), 137-148 https://doi.org/10.18844/cjes.v13i2.2427

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103-127. https://doi.org/10.1037/a0018053

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics - a ‘‘Hopeless’’ issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497-514. https://doi.org/10.1007/BF03173468

Fryer, R. G., & Levitt, S. D., (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, American Economic Association*, *2*(2), 210-240. https://doi.org/10.1257/app.2.2.210

Gallagher, A. M., DeLisi, R., Holst, P. C., McGillicuddy-DeLisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, *75*(3), 165-190. https://doi.org/10.1006/jecp.1999.2532

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, *19*(2), 229-284. https://doi.org/10.1017/S0140525X00042400

Gierl, M., Jodoin, G. M., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large*. Paper Presented at the Annual Meeting of the American Educational Research Association (AERA). New Orleans, Louisiana, USA. Retrieved from http://www.education.ual-berta.ca/educ/psych/crame/

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological science*, *24*(10), 2079-2087. https://doi.org/10.1177/0956797613486989

Halpern, D. F. (2000). S*ex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum https://doi.org/10.4324/9781410605290

Hambleton, R. K., & Rogers, H. J. (1995). Item bias review. *Practical Assessment, Research and Evaluation*, *4*(6), 1-3. Retrieved from http://pareonline.net/getvn.asp?v=4&n=6

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American journal of education*, *112*(2), 209-238. https://doi.org/10.1086/498995

Hox, J. (2002). *Quantitative methodology series. Multilevel analysis techniques and applications*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139-155. https://doi.org/10.1037/0033-2909.107.2.139

Kalaycıoğlu, D. B., & Kelecioğlu, H. (2011). Öğrenci seçme sınavının madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim*, *36*, 3-13. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/143

Kamata, A. (1998). *Some generalizations of the Rasch Model: An application of the Hierarchical Generalized Linear Model* (Doctoral dissertation). Retrieved from https://search.proquest.com/pqdtglobal/docview/304431757/F4C04EDC6C194515PQ/1?accountid=16733

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Kane, J. M., & Mertz, J. E. (2012). Debunking myths about gender and mathematics performance. *Notices of the Ams*, *59*(1). 10-12. https://doi.org/10.1090/noti790

Liu, O. L., & Wilson, M. (2009). 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. https://doi.org/10.1080/15305050902733547

Lubienski, S. T., McGraw, R., & Strutchens, M. E. (2004). NAEP findings regarding gender: Mathematics achievement, student affect and learning practices. In P. Kloosterman, F. K. Lester, & P. A. Kenney (Eds.), *The 1990 to 2000 mathematics assessments of the National Assessment of Educational Progress: Results and interpretations*. Reston, VA: NCTM.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*(4), 316-345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, *16*(10), 404-406. https://doi.org/10.1016/j.tics.2012.06.008

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race and socioeconomic status. *Journal for Research in Mathematics Education*, *37*(2), 129-150. https://doi.org/10.2307/30034845

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*, *44*(5), 351-373. https://doi.org/10.1016/j.jsp.2006.04.004

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*, *17*(4), 297-334. https://doi.org/10.1177/014662169301700401

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9822

OECD. (2012). *PISA 2009 technical report*. PISA, OECD Publishing. https://doi.org/10.1787/9789264167872-en

OECD. (2013). *PISA 2012 released mathematics items*. OECD Publishing. Retrieved from file:///F:/1111/DIf%20anket%20sonuçları/DIF/pisa2012-2006-rel-items-maths-ENG.pdf

OECD. (2014a). *PISA 2012 results: What students know and can do – student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014). PISA, OECD Publishing. https://doi.org/10.1787/9789264201118-en

OECD. (2014b). *PISA 2012 technical report*. PISA, OECD Publishing. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543-578. https://doi.org/10.3102/00346543066004543

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, *24*(2), 124-139. https://doi.org/10.1006/ceps.1998.0991

Penner, A. M. (2003). International gender × item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, *95*(3), 650-655. https://doi.org/10.1037/0022-0663.95.3.650

Qian X. (2011). *A multilevel differential item functioning analysis of trends in international mathematics and science study: Potential sources of gender and minority differences among U.S. eight graders' science achievement* (Unpublished dissertation). University of Delaware, U.S.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268-302. https://doi.org/10.3102/0002831210372249

Rogers H.J. & Swaminathan H. (1993). A comparison of the logistic regression and Mantel–Haenszel procedures for detecting dierential item functioning. *Applied Psychological Measurement*, 17(2), 105-116. https://doi.org/10.1177/014662169301700201

Sanders, J., & Nelson, S. C. (2004). Closing gender gaps in science. *Educational Leadership*, *62*(3), 74-77. Retrieved from http://www.josanders.com/pdf/science.pdf

Siegle, D., & Reis, S. M. (1998). Gender differences in teacher and student perceptions of gifted students' ability and effort. *Gifted Child Quarterly*, *42*(1), 39-47. https://doi.org/10.1177/001698629804200105

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publications.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Turgut, M. F., & Baykul, Y. (2014). *Eğitimde ölçme ve değerlendirme metotları*. Ankara: Pegem Akademi Yayıncılık, Turkey.

Vahedi, S., & Farrokhi, F. (2011). A confirmatory factor analysis of the structure of abbreviated math anxiety scale. *Iranian journal of psychiatry*, *6*(2), 47-53. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3395944/

Vukovic, R. K., Kieffer, M. J., Bailey, S. P., & Harari, R. R. (2013). Mathematics anxiety in young children: Concurrent and longitudinal associations with mathematical performance. *Contemporary Educational Psychology*, *38*(1), 1-10. https://doi.org/10.1016=j.cedpsych.2012.09.001

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., & Blumenfeld, P. C. (1997). Changes in children's competence beliefs and subjective task values across the elementary school years: A three-year study. *Journal of Educational Psychology*, *89*, 451-469. https://doi.org/10.1037/0022-0663.89.3.451

Yalcin, S. (2015). TIMSS 2011 *Fen uygulamasında cinsiyete göre farklılaşan madde fonksiyonunu medde, öğrenci ve okul düzeyinde açıklayan değişkenler* (Doctoral dissertation). Retrieved from http://acikarsiv.ankara.edu.tr/eng/browse/30163/

Zimmerman, B., & Kitsantas, A. (2005). Homework practice and academic achievement. the mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, *30*(4), 397-417. https://doi.org/10.1016/j.cedpsych.2005.05.003

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential ıtem functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

CHARTS

In January, the new CDs of the bands *4U2Rock* and *The Kicking Kangaroos* were released. In February, the CDs of the bands *No One's Darling* and *The Metalfolkies* followed. The following graph shows the sales of the bands' CDs from January to June.

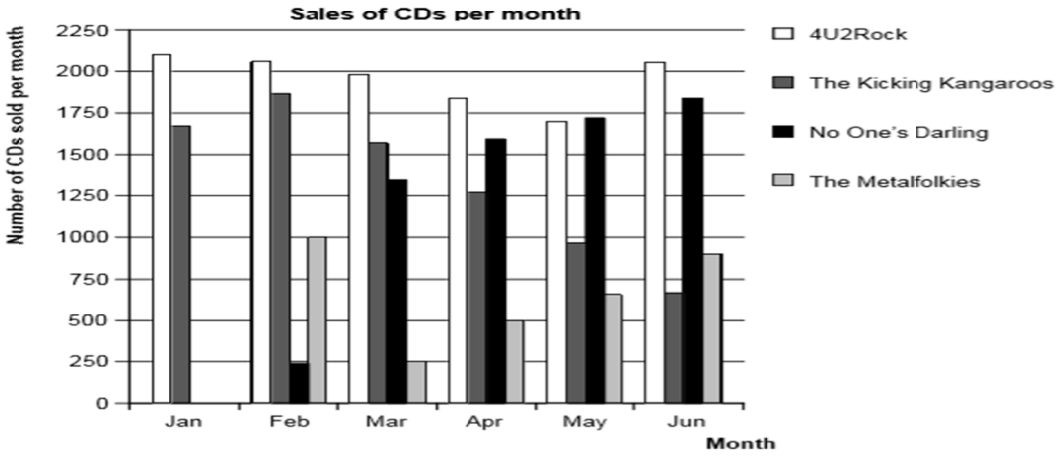

Figure A1. Released sample question in PISA 2012 (OECD, 2013)