Education and Science

Original Article

Vol 49 (2024) No 217 201-223

Examination of the Differential Item Functioning in the PISA 2018 Reading Test Items

Şerife Zeybekoğlu [1], Ayşe Bilicioğlu Güneş [2], Evrim Yalçın [3]

Abstract

This study aimed to conduct Differential Item Functioning (DIF) determination studies using different methods on items in the Program for International Student Assessment (PISA) 2018 reading test in Turkey and compare the performance of the methods used. In the analyses, considering the individualized test design, item packages in the same test(s) in the core, first-stage, and second-stage sections were used. The second package (Core RC2), second package (Stage 1-R12H), and third package (Stage 2-R23H) were selected for the core, first, and second stages, respectively. Three partially scored items in this package were excluded from the analysis, and 33 common items with score of 1-0 were included in the analysis. This study included 147 Turkish students who responded to package of items. The variables of gender (ST004D01T), school location (SC001Q01TA) and index of economic, cultural and social status (ESCS) were drawn from student- and school-scale data and combined with the items of the cognitive test. Prior to data analysis, the dataset was organized, missing data and outliers were examined, and the assumptions of the theories were tested. Within the scope of the study, the Mantel-Haenszel (MH), Logistic Regression (LR), SIBTEST, and Raju's Area Measures methods were employed for two categorical variables, and the Generalized MH, Generalized LR, and Generalized Lord's χ^2 methods were used for three categorical variables. According to gender variables, two, four, and three items were found to show DIF in the MH, SIBTEST, and LR methods, respectively, whereas 17 items were found to display DIF according to the unsigned area test, and seven items were found to display DIF according to the signed area test in Raju's Area Measures. According to the ESCS variable, two and one items manifested DIF in MH and LR, respectively, while 15 items were found to manifest DIF according to the unsigned area test and eight items manifested DIF according to the signed area test in Raju's Area Measures. None of the items showed DIF when using the SIBTEST method. According to the school location variable, one,

Keywords

PISA 2018

Reading skills

Differential Item Functioning (DIF)

Gender

Index of Economic, Cultural and Social Status (ESCS)

School location

Article Info

Received: 08.16.2022

Accepted: 05.25.2023

Published Online: 11.03.2023

DOI: 10.15390/EB.2023.12123

[1] Ankara University, Faculty of Education, Department of Educational Sciences, Türkiye, serifezeybekoglu79@gmail.com

[2] TED University, Faculty of Education, Department of Educational Sciences, Türkiye, ayse.bilicioglu@tedu.edu.tr

[3] Ankara University, Faculty of Education, Department of Educational Sciences, Türkiye, evrim0626@gmail.com

two and 28 item were found to show DIF in Generalized MH, Generalized LR and Generalized Lord's χ^2 method, respectively. The results of the study indicate that although the Classical Test Theory (CTT) -based- and Item Response Theory (IRT)- based DIF methods are compatible, they differ in the level of DIF. IRT- based methods detect more DIF items than CTT- based methods. Additionally, similar results were obtained using the Generalized MH and LR methods.

Introduction

Reading skills have been important assets from the past to the present. Reading skills, which are present in every stage of education and training, have begun to evolve with technological developments. In other words, the quality and content of reading skills required in recent years have changed. Currently, reading is not only restricted to written sources but also linked with electronic texts. This situation brought about the concept of literacy, which includes being able to use different sources, determining one's direction in the face of an uncertain situation, and understanding the differences between perception and reality (Ministry of National Education [MoNE], 2019). The concept of literacy has also become part of the assessment and is the most prominent feature of the Program for International Student Assessment (PISA), a large-scale test in which Turkey regularly participates. By participating in international assessments, such as PISA, countries have the opportunity to determine their statements by comparing their abilities with those of other countries and reviewing their own education systems.

PISA is a study that focuses on mathematical and science literacy and reading, and periodically determines one of these areas as the main subject every three years. The main subject of the PISA 2018 was reading. In this context, reading skills were defined as follows in PISA 2018 (OECD, 2019): *"Reading skills are the ability to understand, use, evaluate, relate and reflect on texts presented in a variety of ways in order to achieve one's goals, develop the knowledge and potential, and participate in society."*

Students' success in reading is also deemed important in terms of their skills in other academic fields. It is thought that if a student's performance in reading skills is low, he/she will have difficulty acquiring other skills in general (Geske & Ozola, 2008). For individuals to be successful at science and mathematics, they must first read and understand the text and symbols well and interpret what they read. Reading comprehension skills are essential in this respect. While Rindermann and Baumeister (2015) emphasized that it is very important to assess reading performance in the interpretation of student achievement (including science and mathematics performance) in PISA; Akbaşlı, Şahin, and Yaykıran (2016) concluded that reading comprehension is an important predictor of mathematics and science achievement from PISA 2006, 2009 and 2012 data as well as related studies conducted with students and teachers. Fuentes (1998) argued that mathematics and reading go hand-in-hand; in other words, he emphasized that students' reading skills should be improved to increase their mathematics achievement. It is also possible to come across many other studies that reveal the relationship between reading skills and mathematics achievement (Ding & Homer, 2020; Erdem, 2016; Grimm, 2008; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2005; Osterholm, 2005).

In the PISA, measurement tools are prepared in two languages, English and French, and sent to countries participating in tests for translation into their national languages. When tests and questionnaires in the PISA are translated into different languages, it is necessary to ensure equivalence between forms. If a study aims to compare individuals from different linguistic and cultural backgrounds, measurement equivalence must be ensured for the comparison to be meaningful. In international administration, individuals differ in terms of their characteristics. Different test

languages, genders, economic, cultural and social status interfere with test performance. This administration, which is applied internationally and addresses the educational systems of various countries, should contain a minimum level of error. Therefore, for comparisons between countries to be meaningful, the measured construct should be independent of the sample. Sample independence implies that the scores of individuals at the same ability level in different subgroups are equal (Osterlind, 1983). If the condition of group independence cannot be met in the PISA, it cannot be determined whether the resulting differences are attributable to genuine discrepancies or whether they are (are they) supposed to change substantially from one group to another. Therefore, it is important to evaluate these issues in PISA conducted internationally in different languages and cultures because errors in the results may cause validity issues and item bias (Gök, Kabasakal, & Kelecioğlu, 2014).

The fact that the characteristics measured by a measurement tool are not invariant for different groups corresponds to the concept of bias, which is defined as the systematic error in test scores for a certain group (Camilli & Shepard, 1994). If a measurement tool is biased, the validity of the decisions, interpretations, and comparisons made with the data obtained from the measurement tool becomes questionable. Bias is one of the biggest threats to the validity of the results obtained using a measurement tool (Clauser & Mazor, 1998; Kristjansson, Aylesworth, Mcdowell, & Zumbo, 2005; Zumbo, 1999). If an item is biased, giving the correct answer depends on belonging to a group rather than the ability being measured (Osterlind, 1983). In this context, for a test or test item to fulfill the validity requirement, an important criterion is that the item does not have bias (Camilli & Shepard, 1994).

Bias studies involve a process that begins with the use of Differential Item Functioning (DIF) determination methods, and continues with expert opinions. DIF determination methods were used to determine the statistical significance of the bias. Then, expert opinions were consulted to determine whether the source of the differential functioning in the item or items showing DIF was due to the real difference between the groups, called the item impact, or bias.

In the literature, many methods can be used to determine the DIF. DIF determination methods differ according to the use of dichotomous or polytomous variables, and the presence of two or more groups. However, it is possible to classify the methods into two types based on Classical Test Theory (CTT) and Item Response Theory (IRT). The most common CTT-based DIF determination methods are Analysis of Variance, Mantel-Haenszel (MH), SIBTEST, Transformed Item Difficulty, Logistic Regression (LR). Some of the IRT- based methods can be ordered as Lord's χ^2, Raju's Area Measures and Likelihood Ratio Test (Camilli & Shepard, 1994; Hambleton, Swaminathan, & Rogers, 1991). Some IRT-based DIF determination methods rely on item parameter estimates or comparisons of the goodness of fit between item response models and data, whereas others develop statistical tests to measure the difference between the curves obtained from two sets of groups or to test for significance (Thissen, Steinberg, & Wainer, 1993). There are two types of DIF: uniform DIF (UDIF) and nonuniform DIF (NUDIF). In UDIF, the DIF effect is constant across ability levels; in NUDIF, the DIF effect varies in magnitude and direction across ability levels (Camilli & Shepard, 1994). In other words, using IRT terminology, UDIF is represented by parallel item characteristic curves, whereas NUDIF is represented by nonparallel item characteristic curves. To determine UDIF using IRT-based methods, IRT Rasch models or one-parameter logistic models and their extensions are used (Wright & Masters, 1982), whereas to determine NUDIF, two- or three-parameter logistic models and their extensions are used (Hambleton & Swaminathan, 1985).

When the DIF determination studies conducted on large-scale tests in the literature are examined, it is seen that they are mainly based on gender (Acar, 2011; Ateşok Deveci, 2008; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Birjandi & Amini, 2007; Çelik & Özkan, 2020; Hamilton & Snow, 1998; Öğretmen & Doğan, 2004; Zenisky, Hambleton, & Robin, 2004), type of school (Bakan Kalaycıoğlu, 2008; Gök, Kelecioğlu, & Doğan, 2010; Karakaya & Kutlu, 2012; Kelecioğlu, Karabay, & Karabay, 2014; Şenferah, 2015; Yıldırım, 2015), region and culture (Ercikan & Kim, 2005; Kabasakal & Kelecioğlu, 2012; Ulutaş, 2012; Yurdugül & Aşkar, 2004). However, DIF studies conducted at the international level have frequently attempted to determine cultural and linguistic biases (Çet, 2006; Grisay & Monseur, 2007; Gür, 2019; Le, 2006; Sırgancı, 2012; Uzun & Gelbal, 2017). The sample size used in DIF studies, the structure of the data, the way the items are scored, and the methods used may cause changes in the DIF levels of the items in the studies. For this reason, there is a requirement for studies based on the comparison of DIF determination methods to determine which method should be used. Although there are such studies in international literature, the methods used are few, and many methods have not been compared with each other. This study is important because it was conducted on data obtained from the PISA 2018 reading test, which is based on education and training, the use of multiple DIF determination methods based on different foundations, providing an opportunity to compare these methods and to determine the validity of measurement tools administered at the international level. In addition, the inclusion of the gender variable, which is frequently encountered as a source of bias, and the variables of school location (Yurdugül, 2003), which are examined in relatively limited studies and socioeconomic status (Berberoğlu, 1995; Walzebug, 2014) is also important in terms of filling the gap in the literature.

This study aimed to conduct DIF determination studies using different methods on the items in the PISA 2018 reading test for a Turkish sample and compare the performance of the methods used. For this purpose, we examined the DIF according to gender, ESCS, and school location.

Within the scope of the study, it was planned to use MH, LR, SIBTEST and Raju's Area Measures methods for two categorical variables; and Generalized MH, Generalized LR and Generalized Lord's χ^2 methods for three categorical variables.

The sub-problems established in pursuit of the aim of this study is as follows:

1. Do the items in the PISA 2018 reading test display DIF according to gender when analyzed using the MH, LR, SIBTEST, and Raju's Area Measures methods?
2. Do the items in the PISA 2018 reading test display DIF according to the ESCS variable when analyzed using MH, LR, SIBTEST, and Raju's Area Measures methods?
3. Do the items in the PISA 2018 reading test display DIF according to the school location variable when analyzed using Generalized MH, Generalized LR and Generalized Lord's χ^2 methods?

Method

Research Model

Within the scope of this study, we examined whether the items in the PISA 2018 reading test for the Turkish sample showed DIF according to gender, ESCS, and school location. Therefore, it has descriptive research properties. In this study, the items in the PISA 2018 reading test were subjected to DIF analysis using different methods, and the results are described. Descriptive research provides a complete and elaborate description of the current situation (Büyüköztürk, Çakmak, Akgün, Karadeniz, & Demirel, 2009).

Sample

In PISA, the school sample was determined using a stratified random sampling method. In PISA 2018, the strata used for schools in determining the sample for Turkey were the Nomenclature of Territorial Units for Statistics (NTUS): Level 1, school type, an administrative form of the school, school location, and gender. After the schools were identified, students from these schools who participated in the administration were randomly selected. In Turkey, 186 schools and 6890 students participated in PISA 2018 on behalf of 12 regions, according to CUTS Level 1 (MoNE, 2019).

In the analyses conducted in this study, item packages that were in the same testlet in the core, first stage, and second stage were handled by considering the adaptive test design. The second package (Core RC2), second package (Stage 1–R12H), and third package (Stage 2–R23H) were selected for the core, first, and second stages, respectively. A total of 147 Turkish students who responded to the item package were included in this study. The students included in this study met these criteria.

Studies have shown that the power of DIF determination methods increases as sample size increases (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993). Zieky (1993) stated that there should be at least 200 people in each group and 400 people in total to calculate the DIF statistics. This study uses a Turkish sample. When the number of students answering different item packages in the sample was examined, the indicated sample size could not be reached. Therefore, the study was conducted with 147 students. Belzak (2020) stated that many researchers have failed to detect DIF in small sample sizes because there is little evidence that commonly used methods can detect DIF. They showed that a moderate level of DIF could be detected in sample sizes as low as 50-100 (25-50 in each group) using fewer complex models. Therefore, the small sample size can be considered a limitation of this study.

Data Collection Tools

A total of 79 countries participated in PISA 2018; in 70 of the participating countries, both cognitive tests and questionnaires were computer-based, whereas in the remaining nine countries, paper-and-pencil tests were administered (OECD, 2019). In Turkey, PISA 2018 was administered using a computer-based approach.

The dataset for PISA 2018 was downloaded from the OECD PISA website, and a sample from Turkey was selected from this dataset. The number of reading skills items in the cognitive tests in the PISA 2018 cycle was much higher than that in previous cycles, owing to the multi-stage adaptive testing design. In addition to 245 items, the reading test included 65 fluent sentences.

At PISA 2018, adaptive testing was used to measure student achievement more accurately. In previous administrations of PISA, the location of each item in the booklets was predetermined; namely, the items in these booklets were fixed. However, in the PISA 2018 reading test, a dynamic structure was developed, and the items were determined according to students' responses to the previous items (OECD, 2019). The items in the PISA 2018 reading test, structured in this direction, consisted of three stages: core, first, and second.

Initially, the students answered items in the core stage, composed of 7-10 items. The items at this stage can generally be scored automatically. Based on the number of correct answers at this stage, student achievement is classified as low, medium, or high (OECD, 2019). The items in the core stage were prepared such that there was no significant difference between them in terms of the difficulty level. Items in the first and second stages were prepared on two levels: relatively easy and difficult. First-stage items were set according to students' achievements in the core stage. Second-stage questions were then set according to both core and first-stage achievements. In the standard computer-based assessment, 64 different item packages were determined and administered to 75% of students. In the alternative computer-based assessment, 128 different item packages were determined and administered to 25% of students.

In this study, one of the 64 item packages in the standard computer-based assessment were selected. Three partially scored items in this package were excluded from the analysis, and 33 common items scored 1-0 were included in the analysis. These items are listed in Table 1.

Table 1. Common Items and Units

Common Items	Units
CR545Q02S	Machu Picchu
DR545Q04C	Machu Picchu
CR545Q06S	Machu Picchu
CR545Q07S	Machu Picchu
CR424Q02S	Fair Trade
CR424Q03S	Fair Trade
CR424Q07S	Fair Trade
CR404Q03S	Sleep
CR404Q06S	Sleep
CR404Q07S	Sleep
DR404Q10AC	Sleep
DR404Q10BC	Sleep
CR558Q02S	Microwave Ovens
DR558Q12C	Microwave Ovens
DR558Q04C	Microwave Ovens
CR558Q06S	Microwave Ovens
CR558Q09S	Microwave Ovens
CR437Q01S	Narcissus
DR437Q07C	Narcissus
CR437Q06S	Narcissus
CR543Q01S	Alfred Nobel
CR543Q03S	Alfred Nobel
CR543Q04S	Alfred Nobel
CR543Q09S	Alfred Nobel
CR543Q10S	Alfred Nobel
CR543Q13S	Alfred Nobel
DR543Q15C	Alfred Nobel
DR566Q03C	The Skellig Rocks
CR566Q04S	The Skellig Rocks
CR566Q05S	The Skellig Rocks
CR566Q14S	The Skellig Rocks
CR566Q06S	The Skellig Rocks
DR566Q12C	The Skellig Rocks

The variables of gender (ST004D01T), ESCS, and school location (SC001Q01TA), which were studied in the sub-problems, were taken from the student- and school-scale data and merged with the items of the cognitive test.

The ESCS variable was averaged and categorized by defining those below the average as low ESCS and those above the average as high ESCS.

Although the school location variable had five categories, the sample sizes of some subgroups were very small. Therefore, the categories were merged and reduced to three. Those with a population of less than one hundred thousand were categorized as towns, those with a population between one hundred thousand and one million as cities, and those with a population of more than one million as metropolitan areas.

Data Analysis

The dataset was organized before starting the data analysis. Missing data and outliers were analyzed. The missing data analysis revealed that the missing data in the dataset were not randomly distributed according to Little and Rubin's (2002) classification. At this point, the multiple imputation method, a method for dealing with missing data, was used to eliminate the missing data, and a complete dataset was obtained for each sample. Outliers in the dataset were then analyzed. To identify the univariate outliers, the total scores were analyzed by converting them into standard z-scores. The z-score indicated the number of standard deviations from the mean of the observed variables. If the z-scores value exceeds ±3, the data is considered as an outlier (Tabachnick & Fidell, 2013). No outliers were detected in the univariate outlier analysis. The normality of the distribution was examined after missing data and outlier analysis. The results are presented in Table 2 and Figure 1.

Table 2. Measures of Central Tendency

Mode	Median	Arithmetic Mean	Skewness	Kurtosis
22.00	21.00	20.21	-0.351	-0.244

The mode (22.00), median (21.00) and arithmetic mean (20.21) values of the measures of central tendency were examined, and the skewness coefficient (Sk=-0.351) and kurtosis coefficient (Kr=-0.244) were found in the range between ±1.

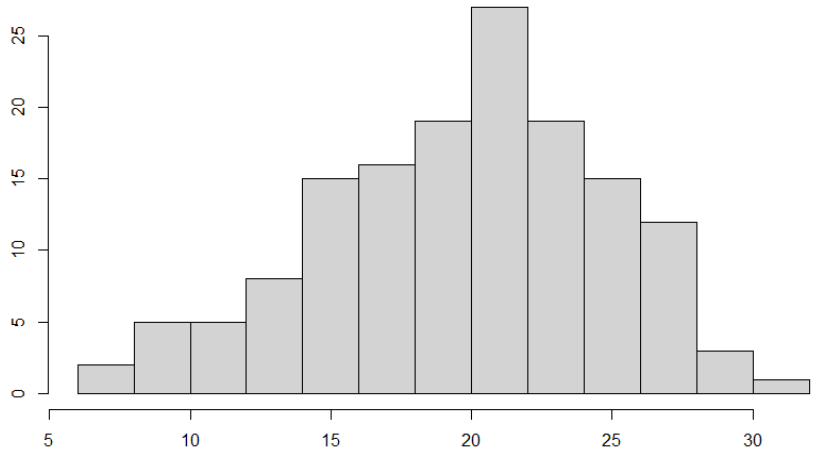

Figure 1. Histogram Chart

Examination of the distribution histogram revealed that the distribution was close to normal. The decision to initiate the study was made with 147 participants. The distribution of these data according to gender, ESCS, and school location is shown in Table 3.

Table 3. Categorical Variables

Gender		ESCS		Location of School		
Female	**Male**	**Low**	**High**	**Town**	**City**	**Metropolitan**
69	77	77	70	45	40	62

In this study, DIF was examined using CTT-based methods (MH, SIBTEST, LR, Generalized MH, and Generalized LR) and IRT-based methods (Raju's Area Measures and Generalized Lord's χ^2). First, the assumptions were examined, and then a DIF analysis was performed. This is because DIF methods based on CTT require the assumption of "unidimensionality," while those based on IRT require the assumptions of "unidimensionality, local independence, and model- data fit."

Analyzing of Assumptions

Unidimensionality

Examining the unidimensionality assumption was carried out with parallel analysis and "Scree Plot" resulting from it is given in Figure 2.

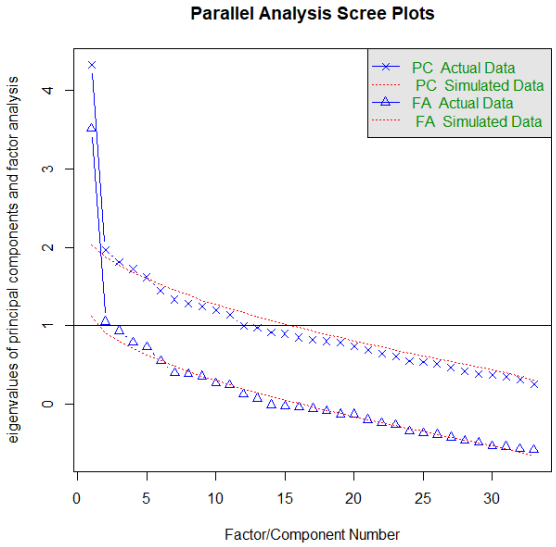

Figure 2. Scree Plot

As a result, it was determined that this assumption was met, and the structure was unidimensional.

Local Independence

The Yen's Q_3 statistic was used to determine the local independence assumption. The results are summarized in Table 4.

Table 4. Results of Yen's Q_3 Statistic

M	SD	Min	%10	%25	%50	%75	%90	Max
-0.030	0.093	-0.312	-0.156	-0.094	-0.024	0.035	0.080	0.293

The Yen's Q_3 statistic shows that the residual correlation is -0.030. Because this correlation is below 0.20, local independence is ensured (DeMars, 2016).

Model- Data Fit

Since the data were dichotomously scored as 1-0, One-Parameter Logistic Model (1PL), Two-Parameter Logistic Model (2PL), and Three-Parameter Logistic Model (3PL) estimations were made. The results are shown in Table 5 and Table 6.

Table 5. ANOVA Results between 1PL and 2PL Model

	AIC	BIC	log.Lik	LRT	df	p
1 PL	5656.70	5758.38	-2794.35			
2 PL	5627.26	5824.63	-2747.63	93.44	32	<0.001

Table 6. ANOVA Results between 1PL and 2PL Model

	AIC	BIC	log.Lik	LRT	df	p
2 PL	5627.26	5824.63	-2747.63			
3 PL	5670.01	5966.06	-2736.00	23.25	33	0.896

In Table 5, the ANOVA between the 1 PL and 2 PL models shows a significant difference between the models and the AIC and log values. The AIC and log.Lik fit indices of the 2PL model were lower, indicating a better model-data fit. In Table 6, the comparison between the two PL and 3 PL models shows that there is no significant difference between the models. In this case, the 2PL model provided a better fit index for the comparison between the dichotomous models.

During the study, DIF analyses were performed using "difR," "ltm," "psych," "sirt" and "ShinyItemAnalysis" packages in the R 4.0.3 program. The cut-off points for the effect sizes of the DIF analyses were as follows:

Mantel-Haenszel (MH)

Zieky (1993) developed a classification system for determining DIF based on the MH method, considering the Δ_{MH} value.

Table 7. Zieky's (1993) Classification System

Ranges	DIF Level	Explanation
$\lvert\Delta_{MH}\rvert < 1.0$	A	Negligible or no DIF
$1.0 \leq \lvert\Delta_{MH}\rvert < 1.5$	B	Moderate DIF
$\lvert\Delta_{MH}\rvert \geq 1.5$	C	High DIF

A positive Δ_{MH} value means that the item displays DIF in favor of the focus group, a negative value means that the item displays DIF in favor of the reference group, and a value equal to zero indicates that the item does not display DIF (De Ayala, 2009).

Generalized MH

This method was developed as an extension of the MH method, as an alternative to polytomous data. While the MH method approaches categories as ordinal, the Generalized MH method approaches them as classified (Wang & Su, 2004). Although the Generalized MH method has many advantages, it does not provide detailed information about the type, effect size, and direction of DIF (Fidalgo & Scalon, 2012).

SIBTEST

For DIF determination based on the SIBTEST method, Roussos and Stout (1996) defined a classification to interpret the β index.

Table 8. The classification system defined by Roussos and Stout (1996) for SIBTEST

Ranges	DIF Level	Explanation
$\lvert\beta\rvert < 0.059$	A	Negligible or no DIF
$0.059 \leq \lvert\beta\rvert < 0.088$	B	Moderate DIF
$\lvert\beta\rvert \geq 0.088$	C	High DIF

As a result of the SIBTEST analysis, a positive β index is interpreted as showing DIF in favor of the reference group, while it is negative when it is in favor of the focus group.

Logistic Regression (LR)

In the LR method, the variables are included in the model in that order. While Model-1 includes the total score, Model 2 also includes the group variables, and Model 3 includes the interaction variable in addition to the group and total scores. With the models composed, it can be determined whether the item shows a UDIF or NUDIF. The difference between the R^2 values of each model yielded the UDIF and NUDIF values. A higher R^2 value was considered when determining UDIF and NUDIF. By comparing the Nagelkerke R^2 values obtained from Model-3 and Model-1, the ΔR^2 value was obtained and the effect size was calculated. The LR classification system was developed by Zumbo and Thomas (1996) and Jodoin and Gierl (2001).

Table 9. The classification system defined by Zumbo & Thomas (1996) and Jodoin & Gierl (2001)

Ranges		DIF Level	Explanation
Zumbo & Thomas	**Jodoin & Gierl**		
$\Delta R^2 < 0.13$	$\Delta R^2 < 0.035$	A	Negligible or no DIF
$0.13 \leq \Delta R^2 < 0.26$	$0.035 \leq \Delta R^2 < 0.070$	B	Moderate DIF
$\Delta R^2 \geq 0.26$	$\Delta R^2 \geq 0.070$	C	High DIF

Generalized LR

Generalized LR is an extension of LR. Using this method, both UDIF and NUDIF estimations can be made for multiple groups (Magis, Raîche, Béland, & Gérard, 2011). In this method, which includes a model with a matching criterion, the statistical significance of the parameters related to group membership and group-score interaction was evaluated using a likelihood-ratio test. If there is a relationship between item responses and group membership, the item manifests a DIF (Magis et al., 2011).

Raju's Area Measures

Signed and unsigned area indices were calculated while analyzing Raju's Area Measures. If these indices are negative, it indicates that DIF exists in favor of the focus group; if they are positive, it indicates that DIF exists in favor of the reference group. To mention the existence of NUDIF, the signed area index value should be less than the unsigned area index value, whereas to mention the existence of UDIF, the signed area index value should be greater than or equal to the unsigned area index value. The greatest disadvantage of Raju's Differential Test and Item Function (DFIT) is the lack of effect size. Oshima and Wright (2015) proposed an approach to define the effect size based on MH in DFIT. With this approach, Δ_{MH} value calculated is considered and divided by the approximate constant K=-15 to obtain β_u (Shealy & Stout, 1993). In the 1 PL and 2 PL models, the constant K is approximately -15, while in the 3 PL models, it is approximately -17.

$$\beta u = \Delta_{MH}/K$$

The Non-compensatory Differential Item Functioning (NCDIF) value was calculated by squaring the obtained β_u value.

$$NCDIF = (\Delta_{MH}/K)^2$$

Oshima and Wright (2015) proposed a classification for the NCDIF, which is presented in Table 10.

Table 10. The classification system defined by Oshima and Wright (2015) for NCDIF

Ranges	DIF Level	Explanation
NCDIF < 0.003	A	Negligible or no DIF
0.003 ≤ NCDIF < 0.008	B	Moderate DIF
NCDIF ≥ 0.008	C	High DIF

Generalized Lord's χ^2

When testing the Lord's χ^2 statistic, item parameters and covariance are calculated for subgroups and the estimated parameters are transformed into a common scale. Thus, the Lord's χ^2 statistic can be calculated using the scaled parameter and covariance values. The obtained values were then compared with the cut-off value in the chi-square test table according to the degrees of freedom to examine the presence of DIF (Camilli & Shepard, 1994). Kim, Cohen, and Park (1995) developed the Generalized Lord's χ^2 method by generalizing Lord's χ^2 method for use in more than two groups.

Results

Findings Related to the First Sub-Problem

We investigated whether the items in the PISA 2018 reading test displayed DIF according to gender, using the MH, SIBTEST, LR, and Raju's Area Measures.

Table 11 lists the DIF items according to the MH method.

Table 11. MH Results for the Gender Variable

Items	α_{MH}	χ^2	p	Δ_{MH}	DIF Level	In Favor of
DR558Q12C	2.6464	2.4669	0.0136*	-2.2870	C	Reference Group
CR566Q05S	2.5394	2.3016	0.0214*	-2.1900	C	Reference Group

Note. Reference Group: male (N=78), Focus Group: female (N=69); *p<.05

In Table 11, the Δ_{MH} values of the items with significant p-values are examined and the level of DIF is determined by comparing it with the Δ_{MH} threshold classified by Zieky (1993) for the MH method. Accordingly, the two items showed a high-level (C) DIF.

The items showing DIF favored the group according to whether the Δ_{MH} value was positive or negative. Items DR558Q12C and CR566Q05S showed DIF in favor of the reference group (males).

Items showing DIF according to the SIBTEST method are shown in Table 12.

Table 12. SIBTEST Results for the Gender Variable

Items	β	SH	χ^2	p	DIF Level	In Favor of
CR545Q06S	0.5413	0.2365	5.2376	0.0221*	C	Reference Group
CR558Q09S	-0.5417	0.2559	4.4817	0.0343*	C	Focus Group
DR437Q07C	-0.5101	0.2324	4.8169	0.0282*	C	Focus Group
CR566Q14S	0.5958	0.2108	7.9870	0.0047**	C	Reference Group

Note. Reference group: male (N=78), Focus group: female (N=69); *p<.05 **p<.01

In Table 12, the β values of the items with significant p-values are compared with the β index effect size interpretation criteria suggested by Roussos and Stout (1996) and the DIF level is determined. Accordingly, four items showed a high-level (C) DIF.

Based on whether the β value was positive or negative, it was determined that the items manifested DIF favoring each group. Items CR545Q06S and CR566Q14S were found to show DIF in favor of the reference group (male), and items CR558Q09S and DR437Q07C in favor of the focus group (female).

Table 13 lists the items showing DIF according to the LR method.

Table 13. LR Results for the Gender Variable

Items	Uniform DIF R^2	Non-uniform DIF R^2	χ^2	p	R^2	DIF Level (Jodoin & Gierl)	Type of DIF
DR558Q12C	0.0612	0.0089	7.0653	0.0079**	0.0612	B	Uniform
CR566Q05S	0.0297	0.0388	4.8815	0.0271*	0.0388	B	Non-uniform
CR566Q14S	0.0092	0.0497	5.3366	0.0209*	0.0497	B	Non-uniform

Note. Reference Group: male (N=78), Focus Group: female (N=69); *p<.05 **p<.01

In Table 13, the Nagelkerke R^2 values of items with significant p-values are examined, and the classification of Jodoin and Gierl (2001) for the LR method is used to determine the level of DIF. Accordingly, the three items displayed moderate DIF (B).

To determine the DIF type, the magnitudes of UDIF R^2 and NUDIF R^2 were compared. Subsequently, graphical analysis was performed at the item level. Item Characteristic Curves (ICC) for these items are shown in Figure 3.

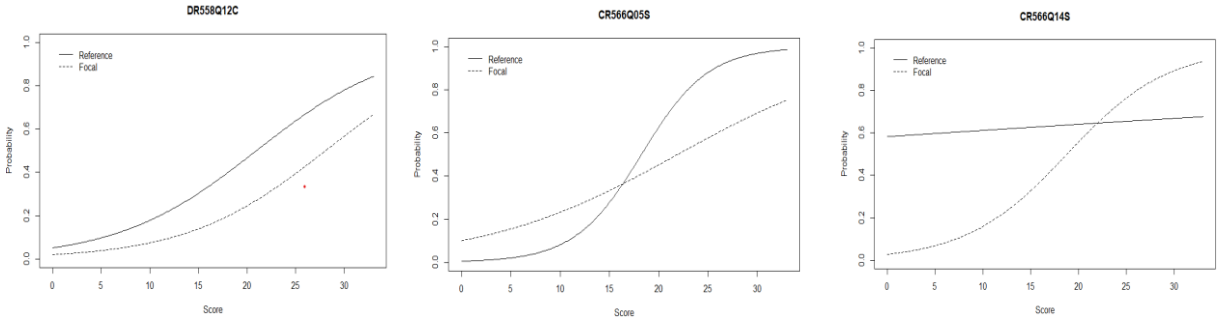

Figure 3. Item Characteristic Curves (Gender)

As a result of the interpretation of ICC, items CR566Q05S and CR566Q14S were found to show NUDIF, and item DR558Q12C was found to show UDIF, which favors the reference group (male).

The items showing the DIF according to Raju's Area Measures method is listed in Table 14.

Table 14. Raju's Area Measures Results for the Gender Variable

Items	Unsigned Area		Signed Area		Δ_{MH}	NCDIF = $(\Delta_{MH}/K)^2$	DIF Level	Type of DIF
	Z	p	Z	p				
DR545Q04C	2.0468	0.0407*	-1.9710	0.0487*	-0.2212	0.000217	A	NU
CR545Q06S	2.6087	0.0091**	2.4553	0.0141*	-0.8812	0.003451	B	NU
CR545Q07S	2.6412	0.0083**	2.2503	0.0244*	-0.6329	0.00178	A	NU
CR404Q06S	2.0720	0.0383*	-0.7329	0.4636	0.1535	0.000105	A	NU
CR404Q07S	2.2147	0.0268*	-2.3455	0.0190*	0.7259	0.002342	A	U
DR404Q10AC	2.9792	0.0029**	-1.2420	0.2142	0.3424	0.000521	A	NU
DR404Q10BC	3.4343	0.0006***	-3.3195	0.0009***	0.5602	0.001395	A	NU
DR558Q04C	2.5646	0.0103*	2.5053	0.0122*	0.6033	0.001618	A	NU
CR558Q06S	2.2707	0.0232*	1.6120	0.1070	0.7825	0.002721	A	NU
CR558Q09S	2.1716	0.0299*	1.3388	0.1806	0.4701	0.000982	A	NU
CR543Q03S	2.8965	0.0038**	2.8860	0.0039**	0.8620	0.003302	B	NU
CR543Q04S	2.0638	0.0390*	1.5308	0.1258	0.9459	0.003977	B	NU
CR543Q09S	2.4450	0.0145*	-1.7340	0.0829	0.6012	0.001606	A	NU
CR543Q13S	2.8596	0.0042**	0.0080	0.9936	0.4772	0.001012	A	NU
CR566Q05S	2.5916	0.0096**	0.3390	0.7346	-2.1900	0.021316	A	NU
CR566Q06S	2.7469	0.0060**	1.7104	0.0872	-0.8469	0.003188	B	NU
DR566Q12C	2.1045	0.0353*	0.7936	0.4274	1.2845	0.007333	B	NU

Note. Reference Group: male (N=78), Focus Group: female (N=69); NU: Non- uniform; U: Uniform; *p<0.5, **p<.01, ***p<.001

The significance of Raju's Z statistic was evaluated as a result of the unsigned area test, and it was determined that all 17 items given in Table 14 displayed DIF in favor of the reference group (male). When the significance of Raju's Z statistic was evaluated as a result of the signed-area test whereas items DR545Q04C, CR404Q07S, and DR404Q10BC showed DIF in favor of the focus group (female), items CR545Q06S, CR545Q07S, DR558Q04C, and CR543Q03S showed DIF in favor of the reference group (male).

In Table 14, the NCDIF values of the items with significant p-values are compared with the effect size interpretation criteria defined by Oshima and Wright (2015) and the DIF level is

determined. Accordingly, it was determined that five items manifested moderate DIF (B) and 12 items manifested negligible DIF (A). In addition, the sign and unsigned area indices were analyzed, and the type of DIF was determined. Accordingly, UDIF was detected in one item and NUDIF in the other 16 items.

The items showing DIF that are commonly detected in MH, LR, and Raju's Area Measures methods according to the gender variable, are given in Table 15.

Table 15. Items Commonly Identified as Showing DIF in Different Methods According to Gender Variable and DIF Levels

	MH	SIBTEST	LR	Raju's Area Measures
DR558Q12C	C		B	
CR566Q05S	C		B	
CR545Q06S		C		B
CR558Q09S		C		A
CR566Q14S		C	B	

Findings Related to the Second Sub-Problem

We investigated whether the items in the PISA 2018 reading test displayed DIF according to the ESCS variables using the MH, SIBTEST, LR, and Raju's Area Measures.

The items that showed DIF according to the results of the MH method are presented in Table 16.

Table 16. MH Results for the ESCS Variable

Items	α_{MH}	χ^2	p	Δ_{MH}	DIF Level	In Favor of
CR543Q10S	0.4536	-1.9635	0.0496*	1.8578	C	Focus Group
CR566Q04S	2.6879	2.4833	0.0130*	-2.3236	C	Reference Group

Note. Reference Group: low (N=77), Focus Group: high (N=70); *p<.05

In Table 16, the Δ_{MH} values of the items with significant p-values are examined and the level of DIF is determined by comparing it with the Δ_{MH} threshold classified by Zieky (1993) for the MH method. Accordingly, the two items showed a high-level (C) DIF.

Based on whether the Δ_{MH} value was positive or negative, it was determined that the items manifested DIF, favoring which group. Item CR543Q10S showed DIF in favor of the focus group (high ESCS), and item CR566Q04S showed DIF in favor of the reference group (low ESCS).

None of the items displayed DIF, according to the SIBTEST method. The items appearing DIF according to the LR method are listed in Table 17.

Table 17. LR Results for the ESCS Variable

Items	Uniform DIF R^2	Non-uniform DIF R^2	χ^2	p	R^2	DIF Level (Jodoin & Gierl)	Type of DIF
CR566Q04S	0.0423	0.0114	4.8156	0.0282*	0.0423	B	Uniform

Note. Reference Group: Low (N=77), Focus Group: High (N=70); *p<.05

In Table 17, the Nagelkerke R^2 values of the items with significant p-values are examined, and the level of DIF items is determined using Jodoin and Gierl's (2001) classification for the LR method. Accordingly, one item exhibited moderate (B) DIF.

To determine the DIF type, the magnitudes of UDIF-R^2 and NUDIF-R^2 were compared. Subsequently, graphical analysis was performed at the item level. The ICC for this item is shown in Figure 4.

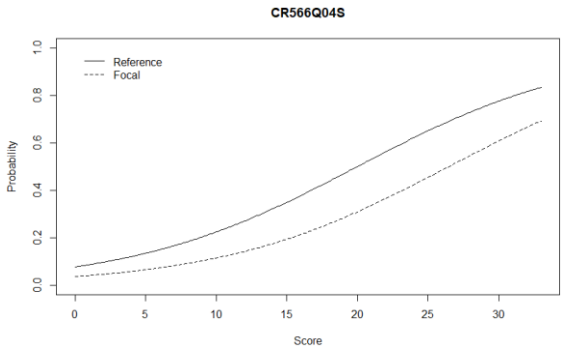

Figure 4. Item Characteristic Curves (ESCS)

As a result of the graphical interpretation, item CR566Q04S was found to reveal UDIF, which was in favor of the reference group (low ESCS).

The items showing the DIF according to Raju's Area Measures method are listed in Table 18.

Table 18. Raju's Area Measures Results for the ESCS Variable

Items	Unsigned Area		Signed Area		Δ_{MH}	NCDIF = $(\Delta_{MH}/K)^2$	DIF Level	Type of DIF
	Z	p	Z	p				
DR545Q04C	-1.7045	0.0883	2.4640	0.0137*	1.3252	0.007805	B	U
CR545Q07S	-2.2176	0.0266*	-1.2392	0.2153	0.3123	0.000433	A	NU
CR404Q06S	-3.0876	0.0020**	2.1162	0.0343*	0.0118	0.000000	A	NU
DR404Q10AC	-5.0506	0.0000***	3.3706	0.0008***	-0.1622	0.000117	A	NU
DR404Q10BC	-5.2041	0.0000***	5.6692	0.0000***	-1.0609	0.005002	B	U
DR558Q04C	-2.4803	0.0131*	-1.9545	0.0506	1.0110	0.004543	B	NU
CR558Q06S	-2.4922	0.0127*	-1.2199	0.2225	-0.0830	0.000031	A	NU
CR437Q01S	-2.5305	0.0114*	2.3427	0.0191*	-1.0418	0.004826	B	NU
CR437Q06S	-2.4266	0.0152*	-0.0066	0.9947	0.2867	0.000365	A	NU
CR543Q04S	-2.1331	0.0329*	-0.9490	0.3426	-1.3023	0.007538	B	NU
CR543Q09S	-5.6812	0.0000***	4.2781	0.0000***	0.2399	0.000256	A	NU
CR543Q13S	-3.5870	0.0003 ***	1.7668	0.0773	-1.3895	0.008581	C	NU
DR543Q15C	-2.4356	0.0149*	2.5932	0.0095**	-0.6247	0.001734	A	U
CR566Q04S	-2.2847	0.0223*	3.4435	0.0006***	-2.3236	0.023996	C	U
CR566Q05S	-2.7794	0.0054**	2.4066	0.0161*	0.6934	0.002137	A	NU
CR566Q06S	-2.7616	0.0058**	-0.3739	0.7085	0.4262	0.000807	A	NU

Note. Reference Group: low (N=77), Focus Group: high (N=70); NU: Non- uniform; U: Uniform; *p<.05

The significance of Raju's Z statistic was evaluated as a result of the unsigned area test, and it was determined that all 15 items given in Table 18 showed DIF in favor of the focus group (high ESCS). When the significance of Raju's Z statistic was evaluated as a result of the signed-area test, eight items displayed DIF in favor of the reference group (low ESCS).

In Table 18, the NCDIF values of the items with significant p-values are compared with the effect size interpretation criteria defined by Oshima and Wright (2015) and the DIF level is determined. Accordingly, two, five, and nine items appeared high (C), moderate (B), and negligible (A) DIF levels, respectively. In addition, the sign and unsigned area indices were analyzed, and the type of DIF was determined. Accordingly, UDIF was detected in four items and NUDIF in the other 12 items.

The items showing DIF commonly detected in MH, LR, and Raju's Area Measures methods according to the ESCS variable are given in Table 19.

Table 19. Items Commonly Identified as Showing DIF in Different Methods According to ESCS Variable and DIF Levels

	MH	SIBTEST	LR	Raju's Area Measures
CR566Q04S	C		B	C

Findings Related to the Third Sub-Problem

It was investigated whether the items in the PISA 2018 reading test displaying DIF according to school location with the method of Generalized MH, Generalized LR and Generalized Lord's χ^2.

Items that manifested DIF according to the Generalized MH method are presented in Table 20.

Table 20. Results of Generalized MH for the School Location Variable

Items	χ^2	p
DR566Q12C	7.6484	0.0218*

Note. Reference Group: metropolitan (N=62), Focus Group (4): city (N=40), Focus Group (3): town (N=45); *p<.05

Table 20 indicates that the p-value of item DR566Q12C was significant and appears the DIF according to the Generalized MH method. Although the Generalized MH method has many advantages, it does not provide detailed information about the type, effect size, or direction of the DIF (Fidalgo & Scalon, 2012).

Table 21 presents the items with DIF according to the Generalized LR method.

Table 21. Generalized LR Results for the School Location Variable

Items	Uniform DIF R^2	Non-uniform DIF R^2	χ^2	p	R^2	DIF Level (Jodoin & Gierl)	Type of DIF
CR404Q06S	0.0453	0.0114	6.0120	0.0495 *	0.0453	B	Uniform
DR566Q12C	0.0821	0.0188	8.6010	0.0136*	0.0821	C	Uniform

Note. Reference Group: metropolitan (N=62), Focus Group (4): city (N=40), Focus Group (3): town (N=45); *p<.05

In Table 21, the Nagelkerke R^2 values of the items with significant p-values are examined, and the level of DIF items is determined using Jodoin and Gierl's (2001) classification for the LR method. Accordingly, one item was found to show moderate DIF (B), and one item was found to show a high DIF (C).

To determine the DIF type, the magnitudes of UDIF- R^2 and NUDIF- R^2 values were compared. Subsequently, graphical analysis was performed at the item level. The ICC for these items are shown in Figure 5.

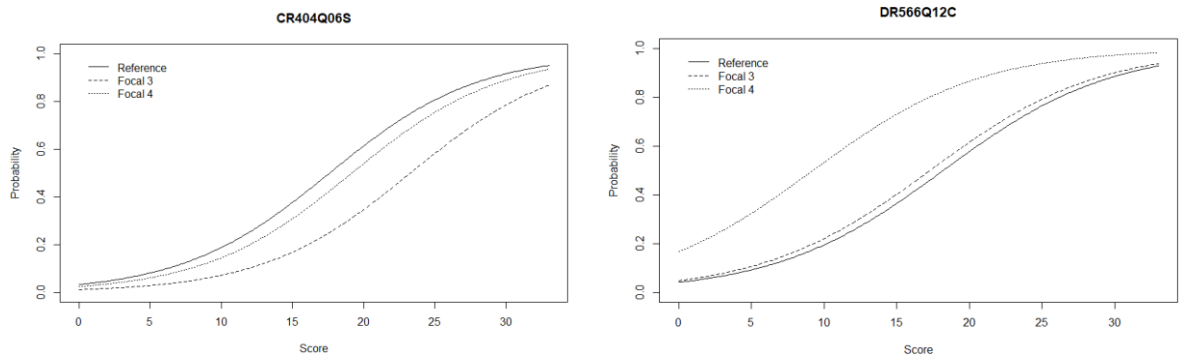

Figure 5. Item Characteristic Curves (School Location)

As a result of the graphical interpretation, items CR404Q06S and DR566Q12C displayed UDIF that favored the reference group (metropolitan) and focus group-4 (city), respectively.

The items showing DIF according to the Generalized Lord's χ^2 method are listed in Table 22.

Table 22. Generalized Lord's χ^2 Method Results for the School Location Variable

Maddeler	χ^2	p
DR545Q04C	53.4426	0.0000 ***
CR545Q06S	78.9994	0.0000 ***
CR545Q07S	44.8603	0.0000 ***
CR424Q03S	15.4787	0.0038 **
CR424Q07S	160.8397	0.0000 ***
CR404Q03S	192.1537	0.0000 ***
CR404Q06S	24.7327	0.0001 ***
CR404Q07S	99.5895	0.0000 ***
DR404Q10AC	15.6739	0.0035 **
DR404Q10BC	48.8002	0.0000 ***
CR558Q02S	48.3320	0.0000 ***
DR558Q12C	17.3317	0.0017 **
DR558Q04C	76.3025	0.0000 ***
CR558Q06S	66.0696	0.0000 ***
CR558Q09S	41.3370	0.0000 ***
CR437Q06S	15.6896	0.0035 **
CR543Q01S	23.2275	0.0001 ***
CR543Q03S	105.8070	0.0000 ***
CR543Q04S	40.0946	0.0000 ***
CR543Q09S	24.0898	0.0001 ***
CR543Q10S	24.4286	0.0001 ***
CR543Q13S	17.4853	0.0016 **
DR543Q15C	209.3311	0.0000 ***
DR566Q03C	74.9375	0.0000 ***
CR566Q05S	16.0128	0.0030 **
CR566Q14S	30.3568	0.0000 ***
CR566Q06S	37.5355	0.0000 ***
DR566Q12C	60.4916	0.0000 ***

Note. Reference Group: metropolitan (N=62), Focus Group (4): city (N=40), Focus Group (3): town (N=45); **p<.01, ***p<.001

Table 22 indicates that the p-value of 22 items was significant, and exhibits the DIF according to the Generalized Lord's χ^2 method.

The items appearing DIF that were commonly detected in the Generalized MH, Generalized LR and Generalized Lord's χ^2 methods according to the ESCS variables are listed in Table 23.

Table 23. Items Commonly Identified as Showing DIF in Different Methods According to School Location Variable and DIF Levels

	Generalized MH	Generalized LR	Generalized Lord's χ^2
DR566Q12C	x	C	x
CR404Q06S		B	x

Discussion, Conclusion, and Suggestions

Within the scope of this study, we aimed to analyze whether the items in the PISA 2018 reading test show DIF according to gender, ESCS, and school location variables using different methods and to compare the results. In the first sub-problem, it was concluded that two items displayed high level (C), four items displayed high level (C), and three items displayed moderate level (B) DIF in the MH, SIBTEST, and LR methods, respectively. As a result of Raju's Area Measures method, while five items manifested moderate DIF (B), 12 items manifested negligible DIF (A). Some DIF items were common in different methods: two items in the MH and LR methods, one item in the LR and SIBTEST methods, and two items in SIBTEST and Raju's Area Measures methods. In the DIF study conducted by Espinoza (2019) in the context of gender for items in the PISA 2015 reading test, most items manifested DIF, albeit at level A. However, it is not surprising that the items reveal DIF in very important large-scale test administrations, such as the PISA. Similar results have been obtained in different DIF studies conducted on gender for PISA as well as internationally administered tests (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017; Lan, 2014; Le, 2009; Lyons-Thomas, Sandilands, & Ercikan, 2014).

In the second sub-problem, two items showed high levels (C) and one item showed moderate levels (B) of DIF in the MH and LR methods, respectively. As a result of Raju's Area Measures method, it was determined that two, five, and nine items appeared high, moderate, and negligible DIF, respectively, and no items showed DIF in the SIBTEST method. When the related items were examined, it was determined that the item coded "CR566Q04S" indicate DIF in all three methods. These results are similar to those of the DIF studies conducted for reading test items by Espinoza (2019) in PISA 2015 and Chen and Jiao (2014) in PISA 2009.

In the third sub-problem, one item in the Generalized LR method revealed a moderate level (B) and one item revealed a high level (C) DIF. According to the Generalized MH and Generalized Lord's χ^2 methods, one and 28 items showed DIF, respectively. When the related items were examined, one item was found to manifest DIF in Generalized LR and Generalized Lord's χ^2 methods, and the item coded "DR566Q12C" was found to exhibit DIF in all three methods. When the literature was examined, similar results were obtained for the Generalized MH and Generalized LR methods when the number of focus groups was more than two. Similar to this study, Uyar and Uyanık (2016), in their DIF analysis using generalized MH and LR methods, found that the methods used yielded approximately the same results.

When these results are considered in general, it can be concluded that the CTT-based- and IRT-based methods are compatible with each other and that the CTT-based methods detected much fewer DIF items than the IRT-based methods. This difference is thought to occur because IRT-based methods are more sensitive and detect even the smallest difference as a DIF. Mazor, Clauser, and Hambleton (1994) stated that IRT-based methods are sensitive to NUDIF, but require a large sample

size and complex calculations. Similarly, Atalay, Gök, Kelecioğlu, and Arsan (2012) concluded that latent score methods are more sensitive and effective than observed score methods in determining DIF items. Pektaş (2018) determined that there was a difference between the CTT- and IRT-based methods in determining DIF items; more DIF items were predicted in the IRT-based methods. The main reason for the different results obtained from the methods used to determine the DIF is that these methods use different statistical techniques and stages for DIF analysis. DIF decisions are made according to the values obtained in these analyses; however, the results are obtained using different mathematical methods (Uzun & Gelbal, 2017). As seen in this study, although the CTT- and IRT-based methods are consistent with each other, there may be differences in DIF levels.

One item each in the gender and ESCS variables displayed DIF according to both the MH and LR methods. The fact that the MH and LR methods produce more consistent results than the other methods is compatible with the results of Gök et al. (2010) and Yurdugül (2003). However, the DIF levels of these items, which are commonly investigated as gender and ESCS variables, differed. While these items manifested a high-level DIF in the MH method, they manifested moderate DIF in the LR method. The difference in DIF levels may be due to the different value ranges used by the different methods used to classify DIF levels. Although both methods analyze total test scores and have similar components, it is thought that the criteria they use in categorization create a difference.

When the findings were examined, it was determined that the methods differed in their abilities to detect UDIF and NUDIF. The MH method detects UDIF, whereas the LR method detects NUDIF (Swaminathan & Rogers, 1990). Swaminathan and Rogers (1990) Rogers and Swaminathan (1993) stated that the LR method is more powerful than the MH method for detecting NUDIF, whereas it is as powerful as the MH method for detecting UDIF. Similarly, Narayanan and Swaminathan (1996) concluded that the SIBTEST and LR methods were more powerful than the MH method for detecting NUDIF. Clauser and Mazor (1998) stated that the results obtained with the SIBTEST method are comparable to those of the MH method; however, because SIBTEST detects NUDIF, it was used to control the findings obtained with the LR method.

A literature review reveals that many studies have compared CTT- and IRT-based DIF determination methods. Item bias studies conducted within the scope of CTT have been criticized for considering some disadvantages of the theory (Shepard, Camilli, & Williams, 1985). These criticisms are that DIF analyses may lead to misinterpretations due to the fact that item parameters change from one group to another, are not constant, or are sample-dependent. It has been argued that IRT-based methods are superior to CTT-based methods (Hambleton & Swaminathan 1985; Lord, Novick, & Birnbaum, 1968). The idea that item parameters do not change from one group to another, that is, they are constant in IRT-based methods; therefore, they provide more accurate results in deciding whether the item is biased or not in comparing groups, has become widespread despite the lack of conclusive evidence. In line with this view, Clauser and Mazor (1998) stated in their study that IRT-based methods are the most widely followed. However, a common problem with IRT-based methods is that meeting the assumptions of the theory and ensuring model-data fit may affect the reliability of the predictions. In this study, the sample size was determined to be a limitation based on IRT assumptions. However, Belzak (2020) showed that when the sample size was small, DIF detection was more accurate in less-complex models.

To sum up the results, the items scored as 1-0 in the PISA 2018 reading test are questionable in terms of item bias. The large number of DIF items prevents accurate interpretation in comparison with these tests. Therefore, the fact that items in the PISA tests seem to provide advantages in favor of some groups makes the validity and reliability of the interpretations made with these tests arguable. However, it should be noted that not every DIF item may manifest bias. This difference may also result from the item impact. Therefore, DIF items should be examined in detail by experts, and how they work in groups should be determined. In addition, since generalizations can be made when these studies are conducted in other countries and similar results are obtained, it may be recommended to conduct DIF studies regularly in international tests such as the PISA.

References

Acar, T. (2011). Sample size in differential item functioning: An application of hierarchical linear modeling. *Educational Sciences: Theory & Practice, 11*(1), 284-288.

Akbaşlı, S., Şahin, M., & Yaykıran , Z. (2016). The effect of reading comprehension on the performance in science and mathematics. *Journal of Education and Practice, 7*(16), 108-121.

Atalay, K., Gök, B., Kelecioğlu, H., & Arsan, N. (2012). Değişen Madde Fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: Bir simülasyon çalışması. *Hacettepe University: Journal of Education*, *43*(43), 270-281.

Ateşok Deveci, N. (2008). *Üniversitelerarası kurul yabancı dil sınavının madde yanlılığı bakımından incelenmesi* (Unpublished doctoral dissertation). Ankara University, Ankara.

Bakan Kalaycıoğlu, D. (2008). *Öğrenci seçme sınavının madde yanlılığı açısından incelenmesi* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Öğrenci seçme sınavının madde yanlılığı açısından incelenmesi. *Education and Science, 36*(161), 3-13.

Belzak, W. C. (2020). Testing Differential Item Functioning in small samples. *Multivariate Behavioral Research*, *55*(5), 722-747.

Berberoğlu, G. (1995). Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. *Studies in Educational Evaluation*, *21*(4), 439-455.

Birjandi, P., & Amini, M. (2007). Differential item functioning (test bias) analysis paradigm across manifest and latent examinee groups (on the construct validity of IELTS). *Journal of Human Sciences*, *8*(2), 1-20.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2009). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.

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

Chen, Y. F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment, 19*(2), 77-96.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Çelik, M., & Özkan, Y. Ö. (2020). Analysis of differential item functioning of PISA 2015 mathematics subtest subject to gender and statistical regions. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(3), 283-301.

Çet, S. (2006). *PISA 2003 matematik maddeleri kullanılarak yanlı çalışan maddelerin tespitinde çok boyutlu eşleştirme analizi* (Unpublished doctoral dissertation). Middle East Technical University, Ankara.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

DeMars, C. (2016). *Item response theory.* Oxford: Oxford University Press.

Ding, H., & Homer, M. (2020). Interpreting mathematics performance in PISA: Taking account of reading performance. *International Journal of Educational Research, 102*, 101566.

Ercikan, K., & Kim, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23-35.

Erdem, E. (2016). Relationship between mathematical reasoning and reading comprehension: The case of the 8th grade. *Necatibey Faculty of Education Electronic Journal of Science & Mathematics Education, 10*(1), 393-414.

Espinoza, J. C. (2019). *Differential item functioning analysis of PISA 2015 reading items: Singapore, Australia, and USA* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

Fidalgo, Á. M., & Scalon, J. D. (2012). Using Mantel-Haenszel methods for detecting differential item functioning. *Psicologia, Reflexão e Crítica, 25*(1), 60.

Fuentes, P. (1998). Reading comprehension in mathematics. *The Clearing House, 72*(2), 81-88.

Geske, A., & Ozola, A. (2008). Factors influencing reading literacy at the primary school level. *Problems of Education in the 21st Century, 6*, 71-77.

Gök, B., Kabasakal, K. A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(1), 72-87.

Gök, B., Kalecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonu belirlemede Mantel- Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Education and Science, 35*(156), 3-16.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*, 410-426. doi:10.1080/87565640801982486

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69-86.

Gür, E. (2019). *PISA 2015 uygulamasındaki maddelerin kültüre göre değişen madde fonksiyonu açısından incelenmesi* (Unpublished master's thesis). Hacettepe University, Ankara.

Hambleton, R. K., & Swaminthan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests* (CSE Technical Report No. 483). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

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*(4), 329-349.

Kabasakal, K. A., & Kelecioğlu, H. (2012). Evaluation of attitude items in PISA 2006 student questionnaire in terms of differential item functioning. *Ankara University Journal of Faculty of Educational Sciences (JFES), 45*(2), 77-96.

Karakaya, İ., & Kutlu, Ö. (2012). Seviye belirleme sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi. *Education and Science, 37*(165), 348-362.

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Seviye Belirleme Sınavı'nın madde yanlılığı açısından incelenmesi. *Elementary Education Online, 13*(3), 934-953.

Kıbrıslıoğlu Uysal, N., & Atalay Kabasakal, K. (2017). The effect of background variables on gender related differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 373-390.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261-276.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953.

Lan, M. C. (2014). *Exploring gender differential item functioning (DIF) in eighth grade mathematics items for the United States and Taiwan* (Doctoral dissertation). University of Washington, Washington.

Le, L. T. (2006, April). *Analysis of differential item functioning*. Paper presented at the Meeting of the American Educational Research Association, San Francisco CA.

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

Lerkkanen, M. K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J. E. (2005). Mathematical performance predicts progress in reading comprehension among 7-year olds. *European Journal of Psychology of Education, 21*(2), 121-137.

Little R. J. A., & Rubin, D. R. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley Publishing Company.

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

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing, 11*(4), 365-386.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*(2), 284-291.

Ministry of National Education. (2019). *PISA 2018 Türkiye ön raporu*. Ankara: MEB.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that non-uniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.

OECD. (2019). *PISA 2018 results volume I: What students know and can do.* Paris: OECD Publishing.

Oshima, T. C., & Wright, K. D. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, *75*(2), 338-358.

Osterholm, M. (2005). Characterizing reading comprehension of mathematical texts. *Educational Studies in Mathematics, 63*, 325-346.

Osterlind, S. J. (1983). *Test item bias*. Thousand Oaks, CA: Sage.

Öğretmen, T., & Doğan, N. (2004). OKÖSYS Matematik alt testine ait maddelerin yanlılık analizi. *Inonu University Journal of the Faculty of Education*, *5*(8), 61-76.

Pektaş, S. (2018). *Değişen madde fonksiyonu belirleme yöntemlerinin test parametreleri kestirimlerine, karar çalışmalarına, g ve phi katsayılarına etkisi* (Unpublished doctoral dissertation). Gazi University, Ankara.

Rindermann, H., & Baumeister, A. E. E. (2015). Parents' SES vs. parental educational behavior and children's development: A re-analysis of the Hart and Risley Study. *Learning and Individual Differences, 37*, 133-138.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355-371.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77-105.

Sırgancı, G. (2012). *PISA 2006 öğrenci anketi madde yanlılığının sıralı lojistik regresyon ve poly-SIBTEST yöntemleri ile test edilmesi* (Unpublished master's thesis. Abant İzzet Baysal University, Bolu.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Şenferah, S. (2015). *2010 Seviye belirleme sınavı matematik alt testi için değişen madde fonksiyonlarının ve madde yanlılığının incelenmesi* (Unpublished master's thesis). Gazi University, Ankara.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-113). Hillsadle, NJ: Lawrence Erlbaum.

Ulutaş, S. (2012). *PISA 2006 fen okuryazarlığı testindeki maddelerin yanlılık bakımından araştırılması* (Unpublished doctoral dissertation). Ankara University, Ankara.

Uyar, S., & Uyanık, G. K. (2016). PISA 2012 Bilişsel maddelerinin kültüre göre değişen madde fonksiyonu bakımından incelenmesi. *Journal of Research in Education and Teaching*, *5*(3), 230-240.

Uzun, N. B., & Gelbal, S. (2017). PISA fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi. *Kastamonu Education Journal*, *25*(6), 2427-2446.

Walzebug, A. (2014). Is there a language-based social disadvantage in solving mathematical items?. *Learning, Culture and Social Interaction*, *3*(2), 159-169.

Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113-144.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. California: MESA Press.

Yıldırım, H. (2015). *2012 yılı seviye belirleme sınavı matematik alt testinin madde yanlılığı açısından incelenmesi* (Unpublished master's thesis. Gazi University, Ankara.

Yurdugül, H. (2003). *Ortaöğretim kurumları seçme ve yerleştirme sınavının madde yanlılığı açısından incelenmesi.* (Unpublished doctoral dissertation). Hacettepe University, Ankara.

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

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, *9*(1-2), 61-78.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item 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.

Zumbo, B. D., & Thomas, D. R. (1996). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia.