ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 128-153

Examination of Differential Item Functioning in PISA 2018 Mathematics Literacy Test with Different Methods

Emre KUCAM* Hamide Deniz GÜLLEROĞLU**

Abstract

This study aims to determine whether the PISA 2018 Mathematical Literacy test items show differential item functioning (DIF) according to gender and parental education level. The sample of the study consisted of a total of 521 students who participated in the practice in Turkey and answered the booklets numbered 1 and 7. The research was conducted on a total of 45 items in these booklets. In this study, the Mantel-Haenszel (MH), Logistic Regression (LR), and Rasch Tree (RT) methods were applied to determine the items showing DIF regarding the gender variable. As a result of the analyses, it was determined that two items in the 1st booklet showed DIF in favour of girls, and an item in the 7th booklet that was common with the 1st booklet showed DIF. This item showed DIF in common for all three methods according to the DIF analyses performed separately by the Mantel Haenszel, Logistic Regression, and Rasch Tree methods. As a result, an item showing DIF in favour of girls was determined with both the MH and LR methods in the 1st and 7th booklets. In addition, when the items in booklets 1 and 7 were examined to see whether they show DIF according to parental education level, it was concluded that an item in booklet 1 was easy for students whose mother's education level was high school, university, and above, but difficult for students whose mother's education level was high school or below.

Keywords: DIF, Logistic Regression, Mantel-Haenszel, PISA, Rasch Tree

Introduction

International research on effective schools and quality research in education regarding developing countries are of great value as sources of information for creating an effective education system (Karip & Köksal, 1996). PISA (Programme for International Student Assessment), which is expressed as the largest international organization that includes all this research, aims to establish the sustainable development of the participating countries with the feedback it gives based on the comparison of the educational statuses of the countries. In this way, a reliable system that is constantly developed, dynamic, effective, and efficient is created. One of the most important stages of this system is the test development. In addition to including important steps to be carried out, the main purpose the test development process is the estimation of validity and reliability. Cronbach (1990) defined the concept of validity as the process of collecting evidence in order to determine the situation of measuring the structure that a measurement tool aims to measure. In line with this, it can be stated that if difficulties are encountered and/or errors are observed in measuring the structure that a measurement tool aims to measure, suspicious situations will arise regarding the quality of the evidence collected. In other words, the error involved in the measurement reduces the validity. If this error is produced systematically and if this error produces results in favour of or against the group/groups taking the item/test, it can be said that this situation creates bias. These results are expressed using two different concepts: test bias and item bias. If the probability of a group answering an item correctly is less than that of another group taking the test due to some characteristics of the item or the test conditions unrelated to the purpose of the test, it is called item bias (Zumbo, 1999). Bias can be defined as a systematic error in test scores depending on a group of individuals (Camilli & Shepard, 1994). To rephrase, in both cases, not all

* PhD student., Ankara University, Faculty of Educational Sciences, Ankara-Turkey, emrekucam@gmail.com , ORCID ID: 0000-0002-4283-7103
** Associate Professor, Ankara University, Faculty of Educational Sciences, Ankara-Turkey, denizgulleroglu@yahoo.com , ORCID ID: 0000-0001-6995-8223

To cite this article:
Kucam, E., & Gülleroğlu, H. D. (2023). Examination of differential item functioning in PISA 2018 mathematics literacy test with different methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 128-153. https://doi.org/10.21031/epod.1122857

Received: 29.05.2022
Accepted: 22.06.2023

individuals taking the item/test are equal on that item/test, which causes the expected measurement results to change against or in favour of a particular group. To reveal this situation, it is important to determine the bias of the measurement tools. While doing this, Differential Item Functioning (DIF) must first be determined by statistical means. DIF is the differentiation of the probability of answering an item correctly among individuals at the same ability level but in a different group. This possible difference should arise from the properties of the items, not from the properties of the subgroups. If an item contains DIF, there is a possibility of bias, however, if an item is biased, it definitely contains DIF. In other words, DIF is necessary but not sufficient for the item bias (Zumbo, 1999). For this reason, it is determined whether an item shows DIF first and then it continues the bias analysis. There are several methods for determining DIF. These methods are summarized below as IRT and CTT-based methods.

Mantel-Haenszel (MH), Simultaneous Item Bias (SIBTEST), and Logistic Regression (LR) methods are examined under the most widely used Classical Test Theory (CTT). On the other hand Lord's Chi-square, Raju's Area Measures and Likelihood-Ratio methods are most used under the Item Response Theory (IRT) (Ellis & Raju, 2003). In the methods examined under CTT and IRT, two types of groups are referred to as reference and focal groups. The focal group is considered to be the disadvantaged group, and the reference group is the group that is advantageous over the focal group. The differentiation of these two groups with respect to each other is determined with statistical methods. DIF is examined under two headings as uniform and non-uniform differential item functioning. Many of the DIF detection methods are designed to reveal the uniform DIF (Jodoin & Gierl, 2001). The uniform DIF is the consistently high level of answering the examined item correctly at all ability levels in a particular group. On the other hand, the non-uniform DIF is the case in which the examined item works in favour of one group in a certain ability level range, while it works in favour of the other group in another ability range (Osterlind & Everson, 2009). Regarding the commonly used DIF determination methods, both Potenza and Dorans (1995) and Alatlı and Şenel (2020) state the theory they are affiliated with, the possibilities of determining uniform or non-uniform DIF, and the number of groups that can be compared in the method as shown in Table 1.

Table 1
Methods of Determining DIF according to Theory, Number of Groups, and Type

Theory	DIF Determination Method	Number of Groups	Uniform/Non-uniform
CTT	Breslow-Day chi-square	2	Non-uniform
	Mantel-Haenszel	2	Uniform
	Simultaneous Item Bias Test-SIBTEST	2	Uniform
	Standardization	2	Uniform
	Transformed Item Difficulties	2	Uniform
	Logistic regression	2	Both
	Generalized logistic regression	>2	Both
	Generalized Mantel Haenszel	>2	Uniform
IRT	Likelihood-Rate Test	2	Both
	Lord's chi-square Test	2	Both
	Raju's Area Measures	2	Both
	Generalized Lord's chi-square Test	>2	Both

Different DIF determination methods have also emerged with the studies conducted after the methods specified in Table 1 were applied. Since the focal and reference groups are predetermined for the methods in the table, these methods are insufficient in determining other potential variables (Zhang, 2009). In addition, the methods in the table focus on only one variable in each implementation, which has the limitation of not being able to focus on the related variables together, especially in large-scale evaluations. The Rasch Tree (RT) method developed for this limitation is one of the new IRT-based methods. The RT method has distinguished among the DIF determination methods because it focuses

on multiple variables together. From this point of view, it can be said that the RT method, which focuses on more than one variable, is more useful than the MH method, which focuses on a single variable. In determining DIF with the Mantel-Haenszel (MH) method, which is one of the methods that deal with only one variable in each application, focal and reference groups are divided into skill or competence layers based on the total test scores. Then, a chi-square probability table is prepared for each skill layer. In the table, the frequencies of correct and incorrect answers are expressed for the groups in each skill layer. The information generated for each skill layer is given in Table 2.

Table 2

Chi-Square Table for Each Skill Layer

Group	Correct Answer	Incorrect Answer	Total
Reference Group	A_j	B_j	n_{Rj}
Focal Group	C_j	D_j	n_{oj}
Total	m_{1j}	m_{0j}	T_j

The ΔMH value is obtained as a result of multiplying the logarithm of the likelihood ratio (αMH) reached with the (∑j Aj Dj / Tj) / (∑j Bj Cj / Tj) operation by -2.35. The DIF levels for these values provided by Zieky (1993) are presented in Table 3.

Table 3

The equivalent of DIF Levels for ΔMH Values

Level of DIF	Condition	Explanation
A	|ΔMH|<1	No or negligible level of DIF
B	1≤|ΔMH|<1.5	Medium Level
C	|ΔMH|≥1.5	High Level

When ΔMH is positive, it is accepted that the items work in favour of the reference group, and when it is negative, it is considered that the items work in favour of the focal group. Another method also used in this study is the Logistic Regression method. Zumbo and Thomas (1997) stated that the 2-degrees-of-freedom chi-square test in the logistic regression should be considered together with the effect size in order to determine DIF. When DIF is determined in large samples without effect size, even insignificant effects may seem statistically significant. In this context, it is recommended to use the ΔR^2 effect size measurement, which is defined as the R^2 difference between the regression models created (Zumbo, 1999). The DIF levels regarding the ΔR^2 effect size values are suggested by Jodoin and Gierl (2001) as follows:

Table 4

The equivalent of DIF Levels for ΔR^2 Values

Level of DIF	Condition	Explanation
A	ΔR^2<.035	No or negligible level of DIF
B	.035≤ ΔR^2 <.070	Medium Level
C	ΔR^2≥.070	Significant Level

When the studies using the MH and LR methods are examined, it is seen that especially large-scale evaluations are studied and different results can be obtained in the same samples (Arslan, 2020; Ayan, 2011; Doğan & Öğretmen, 2008; Gök, Kelecioğlu & Doğan, 2010; Ozarkan, Kucam & Demir, 2017; Schwabe et al., 2014; Şenferah, 2015). It can be said that one of the reasons why different results can be

obtained with the same sample under different methods is the sample size. In these studies, DIF levels are determined based on several variables. The DIF levels are determined according to gender, ethnicity, disability, item type, socioeconomic level, mother tongue, country, content of tests, and affective characteristics (motivation, etc.). Test lengths and sample sizes may also be effective on these variables.

Another method also used in this study is the RT method. In the Rasch model, some of the methods used to determine DIF are for determining DIF in the items, and some are for determining general fit statistics. These methods are designed to compare the parameters of the predefined focal and reference groups. With these methods, it is determined which items may be difficult or easy to answer in which groups, and an opportunity is created to make inferences about what precautions can be taken in these cases. Latent class methods, which have a different understanding from these methods, enable DIF to be determined in groups that have not been defined beforehand and have not been determined to be a possible source of DIF (gender, ethnicity, etc.). Such methods are used in the first stage of the analysis as it is difficult to interpret the groups showing DIF with these methods. Then, the latent classes are tried to be defined. The RT method, on the other hand, combines these two types of DIF determination approaches and reveals a DIF determination method based on the iterative separation technique. In this way, by identifying the groups showing DIF that have not been identified before, direct comments can be made about these groups. It also provides a wide range of opportunities regarding the identification of the DIF sources. The following steps are followed in the RT method (Strobl, Kopf & Zeileis, 2015):

1. First, the item parameters are estimated by including the entire sample.
2. It is statistically tested whether the item parameters differ by considering each covariant.
3. If there are significant instabilities in the covariates at the item parameters, the sample is separated along the covariant with the strongest indecision, and the cut-off point is determined.
4. The process mentioned above is repeated until there is no significant indecision.

In the study of Altıntaş and Kutlu (2019), in which this method was also used, the DIF status according to the country and gender variable was examined by using the data of 615 (Azerbaijan, Bulgaria, and Syria) out of 2476 individuals who took the Ankara University Foreign Student Exam in 2017. In this study, in which the analyses were carried out using the RT method, DIF was determined in 16 items according to the countries. In addition, it was concluded that the exam did not include DIF according to gender. Similarly, the RT method and LR and Rasch methods among the traditional methods were compared regarding the identification of DIF according to gender, ethnicity, socioeconomic level, and mother tongue in Liu's (2017) study with a data set of 731 students studying at the eighth grade of the 2011 TIMSS mathematics subtest in the USA sample. It was determined that 6 items showed DIF in favour of girls with the LR method, 4 items showed DIF in favour of girls and 1 item in favour of boys with the Rasch method, and 2 items showed DIF in favour of girls, and 3 items in favour of boys with the RT method. While 2 of these items for which DIF was determined according to gender were common in all three methods, the results were obtained in favour of girls with the LR method and in favour of boys with other methods in 1 item. In addition, DIF was determined in 7 items related to ethnicity with the RT method. As a result, it was stated that the RT method generated similar results with the LR and Rasch methods in determining the items containing DIF.

Karami, Gramipour, and Minaei (2021), investigated the factors that reveal the differentiation in test items using the Rasch tree method in their study. Data from a special test of the Amin University of Law and Applied Sciences were used to answer the research questions. The data of this simulation study, in which 2414 people participated, were analysed with the DIFtree package in R software, in which the Rasch tree method was used. In the special examination of Amin University of Law and Applied Sciences, it was observed that 9 items showed DIF and most of these items were in the mathematics group, and these items showed DIF according to the age of 18 (second category) and 19 (first category). This study shows that the Rasch tree method is effective in determining the differentiation in test questions.

Asamoah (2020), administered the 10-item, 5-point Likert-type Perceived Stress Scale to 500 participants through a platform called MTurk, which matches practitioners and participants, in his

master's thesis. The data were analysed according to the age, gender, marital status, employment status, social media use, and race variables.According to these data, DIF for gender, ethnic group, employment and social media variables was determined in one item. It was determined that DIF could not be found for the variables of age, marital status and number of children. It was found that the number of items for which DIF was detected by the MH (Mantel-Haenszel) and LR (likelihood ratio test) methods were equal to each other.

In her doctoral thesis, Başman (2017) examined the interactions of the variables of motivation, self-efficacy, and anxiety on the mathematics test items within the scope of changing item function in order to understand the sources of the differences in mathematics achievement of the students participating in the PISA 2012 application. The sample of the research consists of 1084 students who participated in the practice in Turkey. Data were analysed using the Rasch Tree Method (RAY) in the Psychotree package in the R program and the Logistic Regression Likelihood Ratio Method (LROOY) in the Lordif package program. It was determined which items showed DIF according to gender. It was also observed that items showing DIF according to gender determined by RAY showed DIF according to the interaction between gender and intrinsic motivation. It was observed that the DIF status of the items changed both according to a certain threshold value of the girls' intrinsic motivation score and according to the interaction between gender and self-efficacy of mathematics items. It was observed that the DIF status of the items changed according to a certain threshold value of the self-efficacy score of the girls.

In their study, Strobl, Kopf, and Zeileis (2015) suggested the use of the newly named Rasch Tree Method to determine DIF in samples showing DIF but whose group could not be determined beforehand. With this method, DIF in a numerical covariate cannot be overlooked because the numerical covariates (like age) have lots of cutpoints. The exact cutpoint does not need to be pre-specified, the decision is made from the data. This is an advantage of the Rasch tree method.

When all these studies are considered, it is seen that the DIF analyses for large-scale evaluations are mostly made separately on the basis of a single variable and the items containing DIF are determined accordingly. In this case, when the error included in the DIF analysis for each variable in a test is considered, it can be said that the determination of all the variables to be examined whether they are the source of DIF in a single analysis and with a single error will contain statistically fewer errors. In addition, the presence of DIF is the most important threat that may reduce test validity. This type of data obtained from the large-scale exams is thought to be important in terms of identifying the possible sources of DIF.

When the literature is examined, it is seen that there is evidence for the presence of many items showing DIF in the large-scale tests (PISA, TIMSS, PIRLS, etc.) as a result of the analyses made on these tests (Ayan, 2011; Liu, 2017; Schwabe et al., 2014). The presence of the items with DIF even in these applications that fully comply with the test development stages, or more accurately, the presence of items that may constitute bias in these tests arouses suspicion and curiosity about the situation in the national exams prepared without following the test development stages. This is clearly observed in the analyses of the exams held within the scope of the national exams. The methods used are of great importance at the point of questioning the validity of these analyses. In addition, the MH method is frequently used, because it is easy to use and understand, and also because it allows testing the null hypothesis and provides an index showing the size of the DIF (Millsap & Everson, 1993). On the other hand, the LR method can be applied to items that fall into more than one group and ranking scale, and can diagnose regular and irregular DIF (Agresti, 2012). In the RA method, on the other hand, groups showing unidentified DIF can be identified, and direct comments can be made about these groups (Strobl, Kopf & Zeileis, 2015). Therefore, in this study, the DIF level of the items was compared using the LR and RT methods, in addition to the frequently preferred MH method. In this respect, it is expected that this research will contribute to the literature in terms of revealing the weaknesses and strengths of these three methods, determining the items with DIF using these methods in the national exams, and promote studies to be conducted on bias.

In addition, it is seen that DIF determination methods based on CTT and IRT for large-scale evaluations are used extensively in the literature (Altıntaş and Kutlu, 2019; Chen and Thissen, 1997; Doğan and Öğretmen, 2008; Gök, Kelecioğlu and Doğan, 2010), however, the RT method is used relatively less

(Başman, 2017; Liu, 2017; Strobl, Kopf and Zeileis, 2015). In this study, the Rasch Tree method was used, since it handles multiple variables together and the number of subgroups of the parent education level variable is more than 2. It is of great importance to reveal the validity of the measurement tools of PISA, which is one of the large-scale tests, and to realize this with the least amount of error. By comparing the methods based on both the observed score and the IRT, the differences and similarities of the methods were tried to be determined. Since the studies comparing these three methods mentioned above are very few in the literature, the study is important in this respect. The purpose of this research is to determine the differential item functioning (DIF), which varies according to gender and education level, of the PISA 2018 mathematical literacy test items with various methods, in the Turkish sample. For this purpose, the following questions were answered:

1. Do the items in the PISA 2018 mathematics subtest show DIF in the analyses made with the MH, LR, and RT methods according to gender?
2. Do the items in the PISA 2018 mathematics subtest show DIF in the analyses made with the RT method according to the education level of the parents?
3. Are the results regarding DIF coherent in the analyses conducted with the MH, LR and RT methods according to gender?

Method

Research Model

This study aims to determine whether the items in the Turkey sample of the PISA 2018 Mathematical Literacy test show differential item functioning (DIF) according to gender and parental education level and compare the DIF determining methods LR, RT, and MH. In this respect, the research is suitable for the descriptive research as it aims to describe the existing situation. Descriptive research is a research approach that aims to describe a situation as it is (Karasar, 2017).

Population and Sample

The population of the research consists of a total of 521 students, 255 of whom answered booklet number 1 and 266 students who answered booklet number 7 in the PISA 2018 Turkey sample consisting of 6890 people. Booklets 1 and 7 were chosen, because they contain the most common items compared to other booklets. The descriptive statistics regarding the population and sample of the PISA 2018 Turkey application are presented in Table 5.

Table 5
Distribution of PISA 2018 Turkey Population and Sample according to Gender, Class, and School Type

Variable	Group	Population		Sample	
		f	%	f	%
Gender	Boy	3494	50.7	262	50.3
	Girl	3396	49.3	259	49.7
	Total	**6890**	**100**	**521**	**100**
Class	7th Grade	3	0.05	1	0.15
	8th Grade	19	0.3	1	0.15
	9th Grade	1295	18.75	101	19.4
	10th Grade	5360	77.8	401	77
	11th Grade	207	3	17	3.3
	12th Grade	6	0.1	0	0
	Total	**6890**	**100**	**521**	**100**

School Type	Middle school	22	0.2	2	0.3
	General High School (Anatolian High School, Imam Hatip High School, Sports/Fine Arts High School and General High School)	3998	58	307	59
	Science High School	226	3.4	17	3.3
	Social Sciences High School	228	3.4	17	3.3
	Vocational Technical High School	2416	35	178	34.1
	Total	**6890**	**100**	**521**	**100**
Mother's Education Level	Primary School Dropout	7704	10.2	63	12.1
	Primary School	1936	28.1	155	29.8
	Middle School	1519	22	111	21.3
	High School	1079	15.8	86	16.6
	Undergraduate and Above	1580	22.9	100	19.2
	Missing	72	1	6	1
	Total	**890**	**100**	**521**	**100**
Father's Education Level	Primary School Dropout	72	3.9	21	4.1
	Primary School	506	21.8	109	21
	Middle School	1887	27.4	162	31.2
	High School	1492	21.7	100	19.2
	Undergraduate and Above	1653	24	121	23.3
	Missing	80	1.2	8	1.2
	Total	**6890**	**100**	**521**	**100**

Data

In the PISA 2018 Mathematical Literacy test for Turkey sample, 82 items applied in the computer environment were distributed into 36 booklets and used. While preparing the data, twenty-three questions were used in the booklet number 1, and twenty-two questions were used in the booklet number 7. These questions measure mathematical literacy and 11 of the questions in these two booklets are common. The DIF analyses were conducted on these items. The reason for considering booklets 1 and 7 is that the number of common items is the highest compared to other booklets. Dichotomous items were scored as 1-0, while partially scored items were scored as 1 for fully correct answers; it was converted to 0 points for partially correct, incorrect, and blank answers. The 1st and 7th booklets in the PISA 2018 Mathematical Literacy Turkey sample consist of items that are common, partially scored, and scored as dichotomous (1-0). The numbers of common and non-common items selected from these booklets are presented in Table 6.

Table 6

Distribution of Booklets Selected from the PISA 2018 Mathematics Subtest according to Common and Non-Common Items

Booklet Number	Number of Common Items	Number of Non-Common Items	Total
1	12 (3ps*, 9 ds**)	11	23
7	11 (2ps*, 9 ds**)	11	22

**ps: partial scoring*
***ds: dichotomous scoring (1-0)*

When Table 6 is examined, it is seen that 3 of the 12 common items selected from booklets 1 and 7 are scored as partial (ps) and 9 of them are scored as 1-0 (ds). On the other hand, 11 non-common items are scored as 1-0. In addition, the descriptive statistics of the booklets 1 and 7 used in the study are given in Table 7.

Table 7
Descriptive Statistics of Booklets 1 and 7

Descriptive Statistics regarding the Booklets	Booklet 1		Booklet 7	
	Girl	Boy	Girl	Boy
Number of items	23	23	22	22
Number of students	127	1.28	134	132
Mean score	8.18	8.19	9.13	9.05
Median	8	8	8.5	8
Peak Value	8	9	8	5.10
Standard Deviation	4.67	4.47	4.75	4.94
Skewness	.71	.47	.50	.27
Kurtosis	3.09	2.46	2.52	2.05
Lowest score	0	1	2	0
Highest score	21	20	21	21

As presented in Table 7, the mean scores of the girls in the booklets 1 and 7 were 8.18 and 9.13, respectively, while the mean scores of the boys were calculated as 8.19 and 9.05. The fact that the skewness coefficients were positive in both groups indicates that the distribution of scores is slightly skewed to the right. When the distribution of the mean, mode, and median is examined, it is seen that the values are very close to each other, which indicates that the distribution is very close to the normal distribution. When the mean scores of the girls and the boys in the booklets are examined, it can be stated that the values are very close to each other, in other words, the difference in achievement between the girls and the boys in the PISA 2018 Mathematics subtest for booklets 1 and 7 is almost non-existent. Before proceeding to the DIF analysis, the data set was examined in terms of missing values and outliers.

For the DIF analyses to be performed with IRT, it was appointed whether the data obtained from the booklets met the IRT assumptions. These assumptions are unidimensionality, local independence, and model-data fit (Lord, 1980).

The unidimensionality of the mathematical literacy items was examined with the Exploratory Factor Analysis (EFA). For this, the assumptions of EFA were tested first. In this context, the outlier, multivariate normal distribution, linearity, and single-multi-collinearity were examined. However, since it is not possible to directly examine multivariate normality, univariate normality and outliers were examined. The fact that there is not a violation of univariate normality also supports multi-variability (Sharma, 1995).

In this study, Shapiro Wilks test was applied to determine whether the data set meets the normal distribution assumption for booklets 1 and 7 and it was concluded that none of the 23 and 22 items in these booklets respectively showed a normal distribution ($p<.05$). The outliers were obtained by examining the Mahalanobis distances ($p<.001$) and multivariate normality. According to Tabachnick and Fidell (2007), the Mahalanobis Distance value should be compared with the $\chi 2$ table value, which accepts the number of independent variables as the degree of freedom. When the Mahalanobis Distance values are examined, it is seen that there is no value exceeding the critical values of $\chi 2(23)=49.72$ for the 1st booklet and $\chi 2(22)=48.26$ for the 7th booklet. This shows that there is no violation of the outlier and the multivariate normality. When the scatter plots in Figure 1 and Figure 2 are examined, it is seen that the data are clustered on a straight line.

Figure 1

Booklet 1 Mahalanobis Distance Values Scatterplot

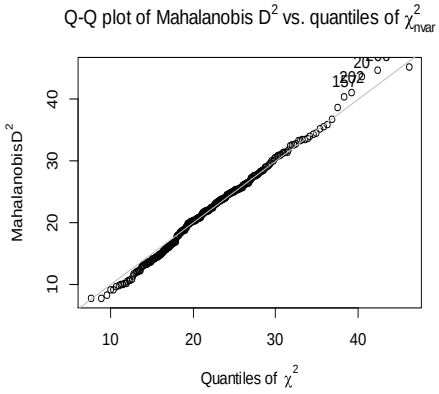
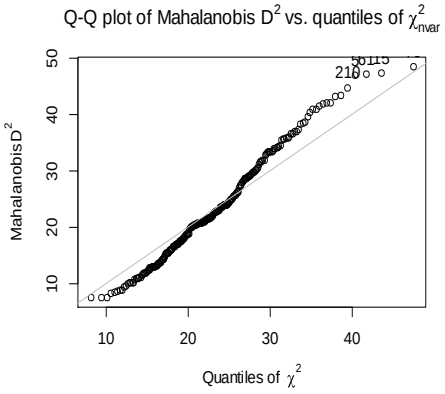


Figure 2

Booklet 7 Mahalanobis Distance Values Scatterplot



Tabachnick and Fidel (2007) stated that the sample size should be 300 or more in order to use factor analytical techniques, but if there is a very strong structure and the representativeness of the group is high, a sample size of up to 150 is acceptable. On the other hand, it is stated by different sources (DeVellis, 2017; Nunnally, 1978; Tavşancıl, 2018) that a sample size of 8-10 times the number of variables/items is sufficient. As the third option, the Kaiser Meyer Olkin (KMO) sample size adequacy test can be applied. In this study, it is noteworthy that the number of students who took the booklets 1 and 7 as the test, namely the sample size, is close to the suggestion of Tabachnick and Fidel (2007) (N1=255, N7=266). (On the other hand, 8-10 times of 23 items makes 184-230, and 8-10 times of 22 items makes 176-220, which shows that this recommendation is more than fulfilled.). As the third option, the KMO test was applied. Since the univariate normal distribution could not be achieved, the KMO test value calculated using the Spearman Rank Differences Correlation matrix was found to be .84 for the 1st booklet and .87 for the 7th booklet. As these values are over .70, it can be stated that the sample size is sufficient for the factor analytical studies.

For the assumption of multicollinearity, the correlation among the variables should be examined and most of the bilateral correlations should be significant (Andy Field, 2012) or Bartlett's Sphericity test can be used. A rough look at the correlation matrix obtained with the Spearman Rank Differences Correlation calculation shows that the pairwise correlations are low but factorable. Bartlett's Test of Sphericity was applied as statistical evidence. As a result, it was determined that the multiple correlations among the variables were statistically significant (Bartlett test of sphericity for the booklet 1; Chi-square=1100.647; df=253 and $p<.05$; Bartlett test of sphericity for the booklet 7; Chi-square=1215.448; df=231 and $p<.05$). In this context, when the correlations among 23 items detected for the booklet 1 in PISA 2018 Mathematical Literacy test are examined, it is seen that the correlations vary between .02 and .26, and when the correlations among 22 items determined for the booklet 7 are examined, it is seen that the correlations vary between .03 and .31. These results indicate that there is no problem of single or multi-collinearity for both booklets.

When multivariate normality, sample size, and the significance of multiple correlations between variables/items were examined, no serious violations were observed that would prevent the use of exploratory factor analysis, provided that the rank-difference coefficient of correlation was used. Thus, it appears that the data met the assumptions of the EFA. In this scope, the EFA was conducted for the 1st booklet (255 people) and the 7th booklet (266 people). Since the univariate normal distribution could not be achieved, the EFA was conducted using the Spearman Rank Correlation Coefficients matrix (Spearman, 1905). For this assumption, it is recommended to examine the eigenvalues and the scree plots of the factors obtained consequently the factor analysis (Cattell, 1966). In this context, the

eigenvalues obtained from the EFA for the 1st and 7th booklet items are shown in Table 8, and the scree plot is presented in Figure 3.

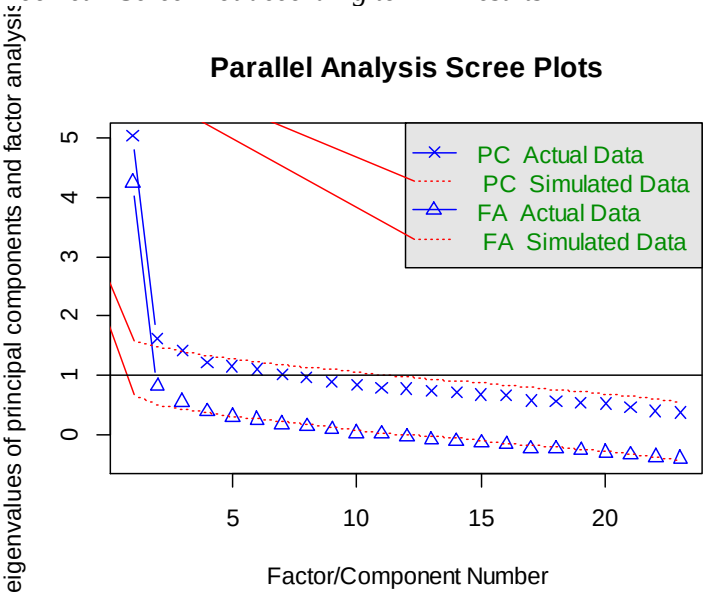
Table 8

Booklet 1 and Booklet 7 Eigenvalues according to EFA Results

Number of factors	Eigenvalues		Variance Explained (%)		Total Variance Explained (%)	
	Booklet 1	Booklet 7	Booklet 1	Booklet 7	Booklet 1	Booklet 7
1	5.03	5.54	22	25	22	25
2	1.63	1.40	7	6	29	31
3	1.41	1.26	6	6	35	37
4	1.21		5		40	

Figure 3

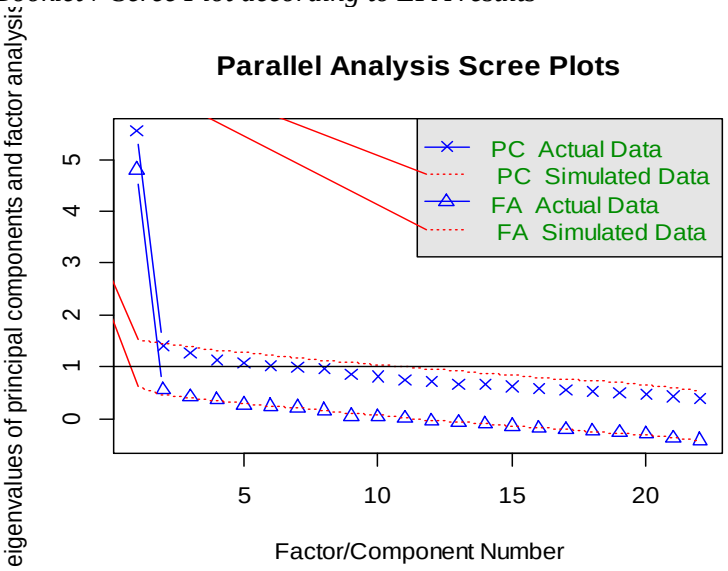
Booklet 1 Scree Plot according to EFA results



When Table 8 is examined, it is seen that the difference in the eigenvalues obtained with the EFA between the first factor of the items in the 1st booklet and the other factors is very large. This shows that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). When Figure 3 is examined, it is seen that a sharp bend is formed with the decrease after the first factor, which indicates that the contributions of the other factors after the first factor to the variance are close to each other and lower than that of the first factor. Local independence, which is one of the other assumptions, is the situation that the answer given to each item is independent of the answers given to the other items (Crocker & Algina, 1986). To test the local independence, Yen's Q3 statistic was calculated. Accordingly, it can be stated that the Q3 values among all the items in the 1st booklet do not exceed .20 (Chen & Thissen, 1997), and thus the local independence assumption is also met.

Figure 4

Booklet 7 Scree Plot according to EFA results



When Table 8 is examined, it is seen that the difference in the eigenvalues obtained with the EFA between the first factor of the items in the 7th booklet and the other factors is very large. This shows that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). When Figure 4 is examined, it is seen that a sharp bend is formed with the decrease after the first factor, which indicates that the contributions of the other factors after the first factor to the variance are close to each other and lower than that of the first factor. Accordingly, it can be stated that the Q3 values among all the items in the 7th booklet do not exceed .20 (Chen & Thissen, 1997), and thus the local independence assumption is also met.

To determine the model-data fit and to carry out the analysis based on IRT, it should be determined which of the 1, 2, and 3 parameter logistic models the data set is compatible with. In the 1st booklet, the Log-Likelihood values obtained for each model and the number of compatible items are presented in Table 10.

Table 9

Booklet 1 Log Likelihood Values of IRT Models and Number of Compatible Items

	1PL	2PL	3PL
Log-Likelihood (LL)	3108.266	3041.361	3022.439
Number of Compatible Items	13	19	17

The fact that the items with a p-value greater than 0.05 are compatible with the model also means the acceptance of the null hypothesis. From this point of view, 13 items are compatible with 1PL, 19 items are compatible with 2PL, and 17 items are compatible with 3PL. The difference between the Log-Likelihood values of the models is taken into account in the evaluation of the model data fit. These difference values are given below:

$$LL_{2PL}-LL_{3PL}=18.922$$

$$LL_{1PL}-LL_{2PL}=60.905$$

LL values showing the chi-square distribution were compared with the critical chi-square value according to the number of items for model-data fit. Since there are 23 items in the 1st booklet, the critical chi-square value is $\chi 2=13.09$, and when compared with the differences above, it is seen that the

difference values are greater than the critical value. In this case, it can be said that the test is compatible with the 3PL model. However, when the number of items compatible with the model is examined, it can be said that the test is coherent with the 2PL model since the number of items compatible with the 2PL model is higher.

In the 7th booklet, the Log-Likelihood values obtained for each model and the number of compatible items are presented in Table 10.

Table 10
Booklet 7 Log-Likelihood Values of IRT Models and Number of Compatible Items

	1PLM	2PLM	3PLM
Log-Likelihood (LL)	2999.812	2955.51	2939.772
Number of Compatible Items	16	21	20

The fact that the items with a p-value greater than 0.05 are compatible with the model also means the acceptance of the null hypothesis. From this point of view, 16 items are compatible with 1PLM, 21 items are compatible with 2PLM, and 20 items are compatible with 3PLM. The difference between the Log Likelihood values of the models is taken into account in the evaluation of the model data fit. These difference values are given below:

$$LL_{2PL}\text{-}LL_{3PL}=15.738$$

$$LL_{1PL}\text{-}LL_{2PL}=44.302$$

LL values showing the chi-square distribution were compared with the critical chi-square value according to the number of items for model-data fit. Since there are 22 items in the 7th booklet, the critical chi-square value is $\chi 2=12.33$, and when compared with the differences above, it is seen that the difference values are greater than the critical value. In this case, it can be said that the test is compatible with the 3PL model. However, when the number of items compatible with the model is examined, it can be said that the test is compatible with the 2PL model since the number of items compatible with the 2PL model is higher. In this case, it can be stated that it is appropriate to choose the 2PLM, in which the majority of the items are compatible, as the IRT model for both booklets.

Data Analysis

To obtain the findings for the first and second research questions, the DIF analyses of the items in the 1st and 7th booklets in the PISA 2018 Mathematics subtest were conducted using the MH, LR, and RT methods. The reference and focal groups required for the analyses were created according to the variables of gender, mother's education level, and father's education level. For MH, the "difMH" command in the "difR" package within the R program was used, and the "raschtree" command in the "psychotree" package within the R program was used for RT. The DIF levels of the items showing DIF for MH and the group in favour of which they showed DIF were determined, and the classification system organized by Zieky (1993) was used for these items.

Results

Findings Regarding Differential Item Functioning According to Gender

Whether the PISA 2018 Mathematics subtest showed DIF according to gender was analysed by the MH, LR, and RT methods, respectively. For this purpose, the items in the 1st booklet and then the ones in the 7th booklet were analysed.

DIF Analysis with Mantel Haenszel Method

The analysis results of the items in the 1st booklet obtained with the MH method are presented in Table 11.

Table 11

Booklet 1 Mantel Haenszel Method Results

Item	Chi-Square	Alpha	Delta	p
CM564Q02S	1.027	0.715	0.785	0.310
CM564Q01S	0.000	1.042	-0.098	0.994
CM571Q01S	0.392	1.283	-0.586	0.530
CM603Q01S	1.200	1.463	-0.895	0.273
DM406Q02C	1.123	0.119	5.003	0.289
DM406Q01C	0.006	0.869	0.329	0.938
CM192Q01S	0.715	1.330	-0.671	0.397
CM423Q01S	0.180	1.263	-0.549	0.671
CM496Q02S	0.055	0.882	0.295	0.814
CM496Q01S	0.402	1.335	-0.679	0.525
CM305Q01S	0.001	0.972	0.064	0.974
CM034Q01S	0.015	0.898	0.250	0.900
DM462Q01C	3.703	0.442	1.916	0.054
CM442Q02S	0.003	0.953	0.112	0.951
CM803Q01S	1.070	0.561	1.356	0.300
CM411Q02S	0.117	0.854	0.370	0.731
CM411Q01S	1.509	0.636	1.060	0.219
CM155Q04S	2.789	1.644	-1.168	0.094
DM155Q03C	4.455	0.331	2.596	0.034*
CM155Q01S	0.015	1.010	-0.023	0.902
DM155Q02C	0.141	0.840	0.407	0.706
CM474Q01S	1.911	1.581	-1.077	0.166
CM033Q01S	0.983	1.382	-0.760	0.321

*p<.05

** Bold item codes refer to the same items in booklets 1 and 7.

When Table 11 is examined, it is seen that only the p-value of the item "DM155Q03C" is significant (p<.05). The ΔMH value of this item was compared with the ΔMH threshold values and it was detected at what level the item showed DIF. Negative values of ΔMH may provide an advantage for the reference group and positive values may provide an advantage for the focal group. In this context, it was determined that the item "DM155Q03C" showed DIF at the C level in favour of the girls forming the focal group. In more general terms, only one of the 5 partially scored items in booklet 1 showed DIF. It is necessary to be careful when generalizing that only one item shows DIF. The finding that female students outperform male students on open-ended items is fitted with this situation (Schwabe et al., 2014; Koğar & Koğar, 2019). The analysis results of the items in the 7th booklet obtained with the MH method are presented in Table 12.

When Table 12 is examined, it is seen that the p-value of none of the items is significant. Negative values of ΔMH may provide an advantage for the reference group, and positive values may provide an advantage for the focal group. However, since negative or positive ΔMH was not significant for any item, it was concluded that none of the items in the 7th booklet showed DIF according to gender.

Table 12

Booklet 7 Mantel Haenszel Method Results

Item	Chi-Square	Alpha	Delta	p
CM034Q01S	0.018	0.989	0.024	0.892
DM462Q01C	1.350	1.616	-1.128	0.245
CM803Q01S	0.346	0.750	0.673	0.555
CM411Q02S	0.000	0.964	0.084	0.982
CM411Q01S	0.019	1.009	-0.022	0.890
CM155Q04S	0.034	0.913	0.212	0.853
DM155Q03C	2.053	1.989	-1.616	0.151
CM155Q01S	0.144	0.861	0.351	0.703
DM155Q02C	0.026	1.117	-0.260	0.871
CM474Q01S	1.362	0.682	0.899	0.243
CM033Q01S	0.796	0.739	0.708	0.372
CM447Q01S	0.000	1.045	-0.105	0.996
CM273Q01S	2.252	1.633	-1.152	0.133
CM408Q01S	0.238	0.807	0.502	0.625
CM420Q01S	0.325	1.246	-0.518	0.568
CM446Q01S	0.184	0.804	0.511	0.668
DM446Q02C	0.768	2.698	-2.333	0.380
CM559Q01S	0.000	0.962	0.090	0.985
DM828Q02C	0.009	0.930	0.170	0.923
CM828Q03S	0.312	1.277	-0.576	0.576
CM464Q01S	0.000	1.078	-0.178	0.982
CM800Q01S	1.311	0.543	1.433	0.252

*p<.05

** Bold item codes refer to the same items in booklets 1 and 7.

DIF Analysis Conducted with Logistic Regression Method

The analysis results of the items in the 1st booklet obtained with the LR method are presented in Table 13.

Table 13

Booklet 1 Logistic Regression Method Results

Item	Chi-Square	R^2	Jodoin&Gierl*	p
CM564Q02S	1.987	0.008	A	0.370
CM564Q01S	1.132	0.005	A	0.567
CM571Q01S	0.419	0.001	A	0.810
CM603Q01S	1.709	0.008	A	0.425
DM406Q02C	7.755	0.080	C	0.020**
DM406Q01C	1.870	0.011	A	0.392
CM192Q01S	6.319	0.026	A	0.042**
CM423Q01S	0.510	0.002	A	0.774
CM496Q02S	1.023	0.003	A	0.599
CM496Q01S	1.489	0.005	A	0.474
CM305Q01S	0.834	0.004	A	0.658
CM034Q01S	0.551	0.002	A	0.758
DM462Q01C	2.952	0.013	A	0.228
CM442Q02S	2.362	0.010	A	0.306
CM803Q01S	1.532	0.008	A	0.464
CM411Q02S	0.889	0.004	A	0.640
CM411Q01S	2.001	0.007	A	0.367
CM155Q04S	3.414	0.016	A	0.181
DM155Q03C	4.822	0.026	A	0.089
CM155Q01S	1.065	0.004	A	0.587
DM155Q02C	0.664	0.002	A	0.717
CM474Q01S	3.084	0.012	A	0,213
CM033Q01S	1.713	0.007	A	0,424

** According to the Jodoin and Gierl effect size, $R^2 \geq 0.070$ means there is a high level (C-level) DIF.*

**Bold item codes refer to the same items in booklets 1 and 7.

When Table 13 is examined, it is seen that only the p-values of the items "DM406Q02C" and "CM192Q01S" are significant (p<.05). The R2 value of these items was compared with the Jodoin and Gierl effect size values and it was determined at what level the items showed DIF.

Figure 5 and Figure 6 present the item characteristic curves of the girls and the boys for these items.

Figure 5
Item Characteristic Curve of the Item DM406Q02C

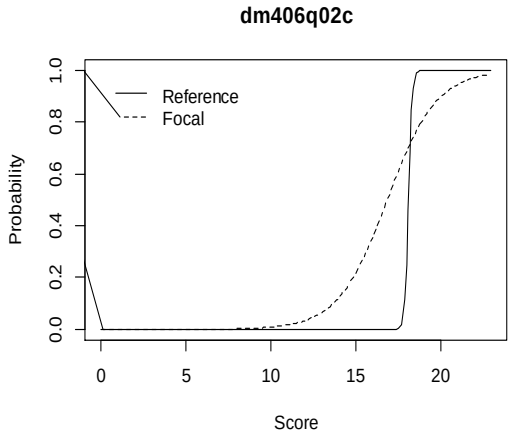
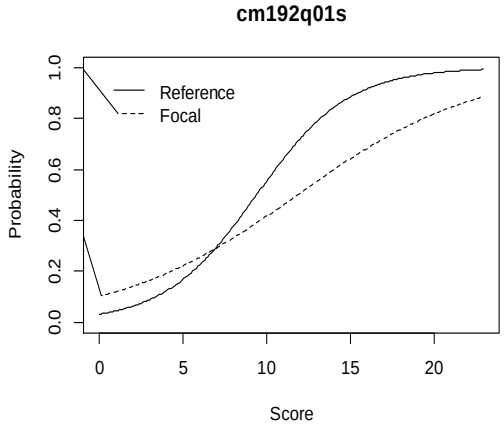


Figure 6
Item Characteristic Curve of the Item CM192Q01S



According to Figure 5, it is seen that the characteristic curve of the item "DM406Q02C" shows DIF at C level in favour of the girls who are in the focal group (R25=.08>.07). When the item characteristic curve is examined, it is seen that the probability of answering the item correctly after 18 points for the reference group boys (reference) increases, and after 10 points for the girls who are the focus group. When a significant DIF is detected for an item, researchers should question whether the DIF actually indicates a bias for the country concerned. That is, it must be decided whether DIF is related to structure (Robitzsch and Lüdtke, 2020). However, since this item cannot be reached, it can be cautiously stated that it is more difficult for men. In Figure 6, on the other hand, it is seen that while the characteristic curve of the item "CM192Q01S" works in favour of the girls (focal) up to about 7 skill levels, it shows DIF in favour of the boys (reference) in the skill group above 7, but the effect size of the DIF likelihood ratio test, which is significant, is at a negligible level (R27=0.026<0.035). The analysis results of the items in the 7th booklet obtained with the LR method are given in Table 14.

Table 14
Booklet 7 Logistic Regression Method Results

Item	Chi-Square	R^2	Jodoin&Gierl*	p
CM034Q01S	2.482	0.009	A	0.289
DM462Q01C	2.233	0.008	A	0.327
CM803Q01S	0.949	0.004	A	0.621
CM411Q02S	2.719	0.012	A	0.256
CM411Q01S	0.887	0.003	A	0.641
CM155Q04S	0.485	0.002	A	0.784
DM155Q03C	7.636	0.044	B	0.022*
CM155Q01S	0.255	0.001	A	0.880
DM155Q02C	0.789	0.002	A	0.674
CM474Q01S	2.515	0.009	A	0.284
CM033Q01S	1.051	0.004	A	0.591

CM447Q01S	1.522	0.005	A	0.467
CM273Q01S	2.796	0.011	A	0.247
CM408Q01S	0.161	0.000	A	0.922
CM420Q01S	0.220	0.000	A	0.895
CM446Q01S	2.711	0.010	A	0.257
DM446Q02C	1.988	0.018	A	0.370
CM559Q01S	0.514	0.001	A	0.773
DM828Q02C	0.182	0.000	A	0.912
CM828Q03S	2.907	0.013	A	0.233
CM464Q01S	0.234	0.000	A	0.889
CM800Q01S	1.910	0.013	A	0.384

***** *According to the Jodoin and Gierl effect size, $R^2 \geq 0.070$ means there is a high level (C-level) DIF.*

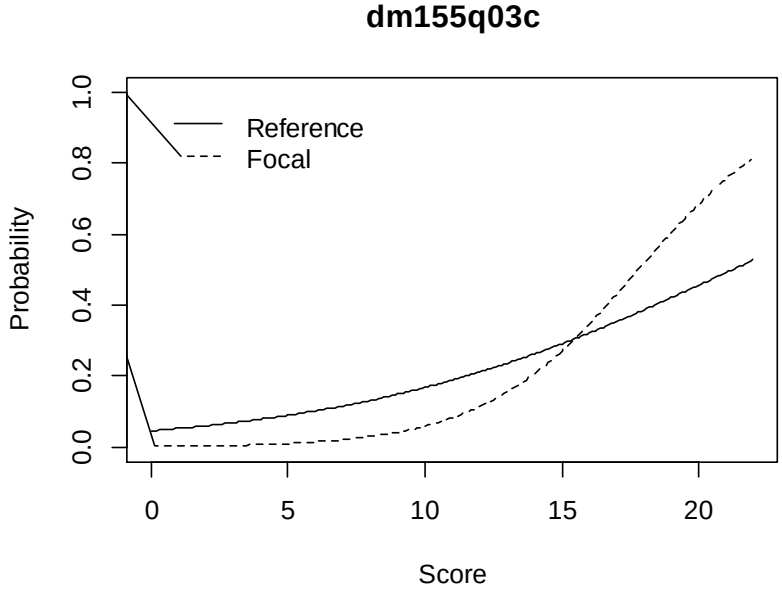
** Bold item codes refer to the same items in booklets 1 and 7.

When Table 14 is examined, it is seen that only the p-value of the item "DM155Q03C" is significant (p<.05). The R2 value of these items was compared with the Jodoin and Gierl effect size values and it was determined at what level the items showed DIF.

Figure 7 presents the item characteristic curves of the girls and the boys for these items.

Figure 7

Item Characteristic Curve of the item DM155Q03C



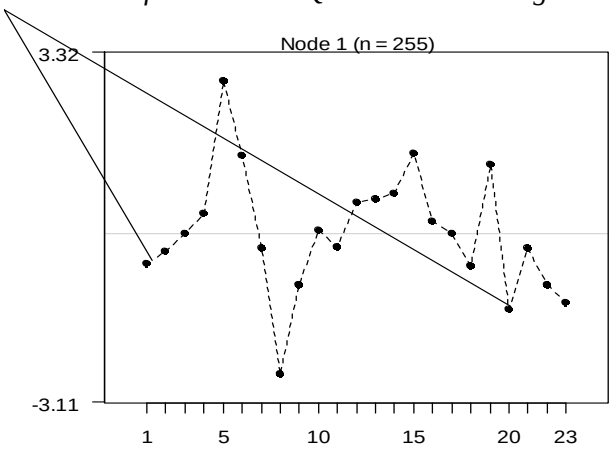
In Figure 7, it was determined that while the item "DM155Q03C" worked in favour of the boys (reference) up to a total score level of approximately 15 in the characteristic curve, it showed non-uniform DIF at the B level in favour of the girls (focal) in the total score group above approximately 15 (R27=0.044>0.035). It is predicted that this is due to the fact that the partially scored items mentioned above are easier for high-achieving girl groups.

DIF Analyses Conducted by Rasch Tree Method

The results of the DIF analysis of the items in the 1st booklet according to gender, obtained with the RT method, are presented in Figure 8.

Figure 8

Rasch Tree for Booklet 1 Questions according to Gender

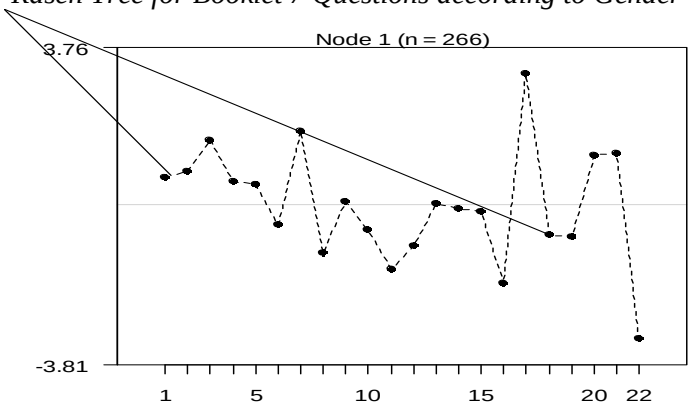


In Figure 8, the items above the horizontal line in the middle are difficult according to the subgroups in the related variable (here, it is gender), while the items on or below the horizontal line in the middle are easy according to the subgroups in the related variable (here, it is gender) (Strobl, Kopf, & Zeileis 2015). However, when Figure 8 is examined, it is inferred that there is no branching according to the subgroups, and 23 items in the 1st booklet, whose item difficulties range from -3.11 to 3.32, do not contain DIF according to gender. Appendix A more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

The DIF analysis results of the items in the 7th Booklet obtained with the RT method according to gender are given in Figure 9.

Figure 9

Rasch Tree for Booklet 7 Questions according to Gender



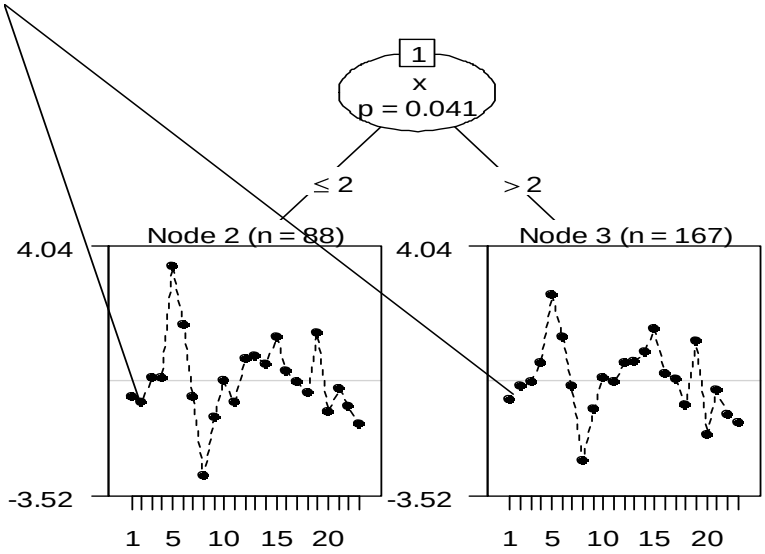
When Figure 9 is examined, it is observed that there is no branching according to the subgroups. In addition, it is determined that 22 items in the 7th booklet, whose item difficulties range from -3.81 to 3.76, do not contain any DIF according to gender. Appendix B more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

Findings Regarding Differential Item Functioning According to Parental Education Level

Whether the PISA 2018 Mathematics subtest shows DIF according to the parental education level was analysed with the RT method. For this purpose, the items in the 1st booklet and then the 7th booklet were analysed. The DIF analysis results of the items in the 1st booklet according to the mother's education level and the father's education level obtained with the RT method are presented in Figure 10.

Figure 10

Rasch Tree for Booklet 1 Questions according to Mother's Education Level

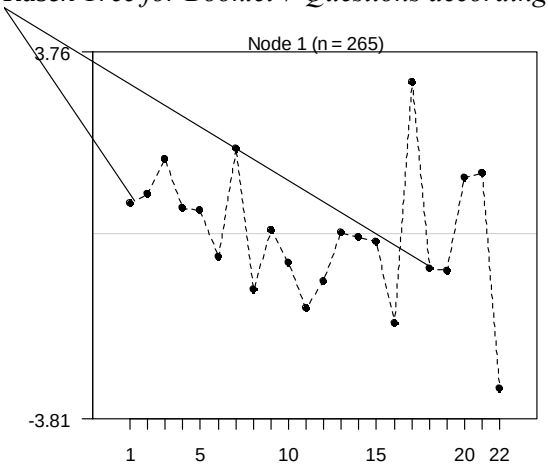


The variable "x" represents the mother's education level, "1" being a university graduate, "2" being a high school graduate, "3" being a secondary school graduate, "4" being a primary school graduate and "5" being a primary school dropout. Strobl, Kopf, and Zeileis (2015) state that item difficulty values below the zero line indicate that the items are easy, while items above the zero line indicate that the items are difficult. According to this statement, some items in PISA 2018 mathematics subtest Booklet 1 show DIF according to the mother's education level.

When Figure 10 is examined, item 11 (CM305Q01S) (ordered points in Figure 2), which is one of 23 items in the 1st booklet, whose item difficulty ranges from -3.52 to 4.04, seems easy for the students whose mother's education level is 2 or below, that is, high school graduate or university graduate, however, it seems more difficult for the students whose mother's education level is above 2, that is, middle school graduate, primary school graduate or primary school dropout. Appendix C more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were. Considering that this item tests visuospatial ability, the significance of the difference between the spatial perceptions of the students whose mother's education level is high school or higher and the spatial perceptions of the students whose mother's education level is secondary school and below coincides with this situation (İrioğlu & Ertekin, 2011). The DIF analysis results of the items in the 7th booklet according to the parental education level obtained with the RT method are presented in Figure 11.

Figure 11

Rasch Tree for Booklet 7 Questions according to Mother's Education Level



In Figure 11, the items above the horizontal line in the middle are difficult according to the subgroups in the related variable, while the items on or below the horizontal line in the middle are easy according to the subgroups in the related variable. However, when Figure 11 is examined, it is seen that there is no branching according to the subgroups and 22 items in the 7th booklet, whose item difficulties range from -3.81 to 3.76, do not contain any DIF according to the parental education level. Appendix D more comprehensively shows what items showed DIF and what exactly the item difficulty parameters were.

Comparison of DIF Analyses Conducted with the MH, LR, and RT Methods According to Gender

The comparison of the DIF analyses according to gender in the PISA 2018 Mathematics subtest in the 1st and 7th booklets is presented in Table 15.

Table 15

Comparison of DIF Analyses according to Gender in Booklets 1 and 7

Booklet Number	Mantel Haenszel (MH)	MH DIF Direction	Logistic Regression (LR)	LR DIF Direction	Rasch Tree (RT)	RT DIF Direction
1	DM155Q03C	Girls	DM406Q02C	Girls	-	-
7	-	-	DM155Q03C	Girls	-	-

When Table 15 is examined, it is seen that DIF was determined in favour of the girls only for the item "DM155Q03C" in the 1st booklet using the MH method. It was also determined by Logistic Regression method that the same item in the 7th booklet (DM155Q03C) contained DIF in favour of the girls. In this respect, MH and LR DIF determination methods are compatible with each other, which is also consistent with the findings of Gök, Kelecioğlu, and Doğan (2010). In the DIF analyses conducted with LR, it was determined that the item "DM406Q02C" in the 1st booklet also contained DIF in favour of the girls. It was noticed that this item also contained DIF at C level in the findings obtained with the MH method, but it was not included in Table 16 because it was not significant. In the DIF analyses conducted with the Rasch Tree method, no DIF was determined for any of the items in the 1st and 7th booklets. This indicates that the RT method differs from the MH and LR methods. Considering the number of items determined to contain DIF, it is seen that the LR method is more sensitive than the RT method, which is in line with the findings of Liu (2017).

Discussion, Conclusion, and Suggestions

In this study, firstly, it was examined whether the items in the 1st and 7th booklets of the PISA 2018 mathematics subtest applied to the Turkish sample showed DIF according to gender. In the DIF analyses conducted with the MH, LR, and RT methods, it was concluded that the items "DM155Q03C" (MH) and "DM406Q02C" (LR) in the 1st booklet showed DIF at the C level in favour of the girls. In the 7th booklet, it was determined that the item "DM155Q03C" (LR), which is common with the 1st booklet, showed DIF at the B level. As a result, the item "DM155Q03C" showed DIF in favour of the girls in both MH and LR methods. It is noteworthy that the items showing DIF are open-ended, that is, partially scored items, regardless of methods applied, which is in line with the findings of Schwabe et al. (2014), Başman (2017), and Koğar and Koğar (2019). In addition, there are studies in the literature showing that the methods based on CTT and IRT are more compatible within themselves (Kan, Sünbül, & Ömür, 2013; Doğan & Öğretmen, 2008). It can be stated that the results obtained from this study are compatible with these studies.

It was examined whether the items in the 1st and 7th booklets of the PISA 2018 mathematics subtest applied to the Turkish sample showed DIF according to parental education level. This analysis was conducted with the RT method since the related variable had more than two categories. In these analyses, the item "CM305Q01S" in the 1st booklet was determined to be easy for the students whose mother's education level is high school graduate or university graduate, however, it is difficult for the students whose mother's education level is below high school level. Considering that the item is visuospatial, this finding coincides with the significant difference between the spatial perceptions of students whose mother's education level is high school or higher and the spatial perceptions of students whose mother's education level is lower than high school (İrioğlu & Ertekin, 2011).

When the literature is examined, it is noteworthy that while it is possible to come across many studies aiming to determine DIF, there are very few studies on determining bias regarding the evaluation of DIF together with the items. In this context, it can be suggested that examining the reasons behind DIF of the items in terms of both the technical and affective properties of the items may be beneficial in terms of increasing the quality of the items. In addition, the items showing DIF according to gender and parental education level were focused on within the scope of this research. However, there are also many different variables such as socioeconomic level and school type, which are thought to affect mathematics achievement. It may also be recommended to conduct studies that examine the underlying causes of the items showing DIF according to these variables.

Declarations

Author Contribution: Emre Kucam-Conceptualization, methodology, analysis, writing & editing, visualization. H.Deniz Gülleroğlu-Conceptualization, methodology, writing-review & editing, supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: PISA data were used in this study. Therefore, ethical approval is not required.

References

Agresti, A. (2012). *Categorical data analysis (Vol. 792)*. John Wiley & Sons. https://doi.org/10.1002/0471249688

Alatlı, B. A., & Şenel, S. (2020). Değişen Madde Fonksiyonunun Belirlenmesinde "difR" R Paketinin Kullanımı: Ortaöğretime Geçiş Sınavı Fen Alt Testi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 1-37. https://doi.org/10.30964/auebfd.684727

Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating Differential Item Functioning of Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model. *International Journal of Assessment Tools in Education, 6*(4), 602-616. https://dx.doi.org/10.21449/ijate.554212

Arslan, M. (2020). Teog Sınavının Yabancı Dil Alt Testine Ait Maddelerin Yanlılığının İncelenmesi. *Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü*.

Asamoah, N. A. B. (2020). *Assessing Differential Item Functioning in the Perceived Stress Scale*. University of Arkansas. https://scholarworks.uark.edu/etd/3775

Ayan, C. (2011). PISA 2009 fen okuryazarlığı alt testinin değişen madde fonksiyonu açısından incelenmesi. *Yayınlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü*.

Başman, M. (2017). Matematik başarısında cinsiyet ve duyuşsal özelliklerin etkileşimine göre Rasch ağacı yöntemi ile değişen madde fonksiyonunun belirlenmesi. *Doktora Tezi, Ankara: Ankara Üniversitesi Eğitim Bilimleri Enstitüsü*.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). Sage. https://doi.org/10.1177/109821409701800108

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22(3), 265-289.* https://doi.org/10.2307/1165285

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. Harper Collins Publishers, New York. https://doi.org/10.1002/sce.3730350432

DeVellis, R. F. (2017). *Ölçek geliştirme kuram ve uygulamalar. (T. Totan, Çev.).* Nobel Akademik Yayıncılık. https://doi.org/10.1177/109821409301400212

Doğan, N., & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede mantel-haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim, 33(148), 100-112.*

Ellis, B. B., & Raju, N. S. (2003). *Differential item and test functioning.* Jossey-Bass/Wiley.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications. https://doi.org/10.1111/insr.12011_21

Gök, B., Kelecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*(156).

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

İrioğlu, Z., & Ertekin, E. (2011). İlköğretim İkinci Kademe Öğrencilerinin Zihinsel Döndürme Becerilerinin Bazı Değişkenler Açısından İncelenmesi. *Journal of Educational and Instructional Studies in the World, 75.*

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. https://doi.org/10.1207/S15324818AME1404_2

Kan, A. ,Sünbül, Ö. ve Ömür, S (2013). 6.-8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 9(2), 207-222.* https://doi.org/10.17860/efd.55452

Karami, H. R., Gramipour, M., & Minaei, A. (2021). *Application of The Rasch Tree Model In The Detection Of Differential Item Functioning* (Case Study: Recruitment Exams Of The Police Of The Islamic Republic Of Iran). https://doi.org/10.22054/jem.2021.61694.2190

Karasar, N. (2017). Bilimsel araştırma yöntemi (2. yazım, 32. Basım). *Nobel Yayın Dağıtım*.

Karip, E., & Köksal, K. (1996). Etkili eğitim sistemlerinin geliştirilmesi. *Kuram ve Uygulamada Egitim Yönetimi Dergisi*, *2*(2), 245-257.

Koğar, E. Y., & Koğar, H. (2019). Investigation of scientific literacy according to different item types: PISA 2015 Turkey sample. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *19*(2), 695-709. https://doi.org/10.17240/aibuefd.2019.19.46660-467271

Liu, M. (2017). *Differential Item Functioning in Large-scale Mathematics Assessments: Comparing the Capabilities of the Rasch Trees Model to Traditional Approaches* (Doctoral dissertation, University of Toledo).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

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

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill. https://doi.org/10.1177/014662169501900308

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications. https://doi.org/10.4135/9781412993913

Ozarkan, H. B., Kucam, E., & Demir, E. (2017). Merkezi ortak sınav matematik alt testinde değişen madde fonksiyonunun görme engeli durumuna göre incelenmesi. *Current Research in Education*, *3*(1), 24-34.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied psychological measurement, 19(1), 23-37.*

Robitzsch, A.; Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psych. Test Assess. Model.* 2020, 62, 233–279. https://bit.ly/3ezBB05 (accessed on 12 June 2023).

Schwabe, F., McElvany, N., Trendtel, M., Gebauer, M. M., & Bos, W. (2014). Vertiefende Analysen zu migrationsbedingten Leistungsdifferenzen in Leseaufgaben. *Zeitschrift für Pädagogische Psychologie*.

Sharma, S. (1995). *Applied multivariate techniques*. John Wiley & Sons, Inc..

Spearman, C. (1905). Proof and disproof of correlation. *The American Journal of Psychology, 16*(2), 228-231.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316. https://doi.org/10.1007/s11336-013-9388-3

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

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Boston, MA: Pearson.

Tavşancıl, E. (2018). *Tutumların ölçülmesi ve SPSS ile veri analizi.* Nobel Akademik Yayıncılık.

Zhang, M. (2009). *Gender related differential item functioning in mathematics tests: A meta-analysis* (Doctoral dissertation, Washington State University).

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development.* Lawrence Erlbaum Associates, Inc.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, 1-57.

APPENDICES

APPENDIX A

Item difficulties parameters for Booklet 1 (for gender)

Item	Female	Male
CM564Q02S	0.39	-0.06
CM564Q01S	0.39	0.02
CM571Q01S	0.70	0.41
CM603Q01S	0.94	0.58
DM406Q02C	4.40	1.17
DM406Q01C	2.27	0.74
CM192Q01S	0.39	0.18
CM423Q01S	-1.72	-0.69
CM496Q02S	-0.04	-0.10
CM496Q01S	0.70	0.48
CM305Q01S	0.58	-0.12
CM034Q01S	1.40	0.56
DM462Q01C	1.61	0.54
CM442Q02S	1.50	0.64
CM803Q01S	2.35	0.65
CM411Q02S	1.12	0.35
CM411Q01S	0.94	0.39
CM155Q04S	-0.01	-0.08
DM155Q03C	2.35	0.66
CM155Q01S	-0.52	-0.41
DM155Q02C	0.58	0.47
CM474Q01S	-0.33	0.003
CM033Q01S	-0.59	-0.21

APPENDIX B

Item difficulty parameters for Booklet 7 (for gender)

Item	Female	Male
CM034Q01S	1.01	0.85
DM462Q01C	1.01	1.07
CM803Q01S	1.84	1.45
CM411Q02S	0.93	0.79
CM411Q01S	0.93	0.69
CM155Q04S	0.11	-0.005
DM155Q03C	1.72	1.87
CM155Q01S	-0.41	-0.51
DM155Q02C	0.60	0.40
CM474Q01S	0.15	-0.21
CM033Q01S	-0.69	-0.86
CM447Q01S	-0.34	-0.32
CM273Q01S	0.32	0.56
CM408Q01S	0.49	0.25
CM420Q01S	0.32	0.28
CM446Q01S	-0.99	-1.08
DM446Q02C	2.97	2.77
CM559Q01S	-0.04	-0.21
DM828Q02C	-0.11	-0.21
CM828Q03S	1.41	1.27
CM464Q01S	1.56	1.23
CM800Q01S	-2.08	-2.14

APPENDIX C

Item difficulty parameters for Booklet 1 (for mother's education level)

Item	Primary School Dropout	Primary School	Middle School	High School	Undergraduate and Above
CM564Q02S	0.003	1.24	0.70	0.53	0.24
CM564Q01S	0.006	1.30	0.73	-0.19	0.54
CM571Q01S	0.006	1.25	0.81	1.09	0.65
CM603Q01S	0.019	1.45	0.81	1.09	0.87
DM406Q02C	0.36	2.32	0.97	3.05	24.88
DM406Q01C	0.02	1.69	0.85	3.05	1.97
CM192Q01S	0.01	1.31	0.70	0.66	0.04
CM423Q01S	-0.02	1.31	0.44	-2.00	-1.79
CM496Q02S	-0.001	1.62	0.63	-0.19	-0.15
CM496Q01S	0.008	1.47	0.72	1.24	0.44
CM305Q01S	0.004	1.41	0.76	0.41	0.13
CM034Q01S	0.01	1.47	0.84	2.02	1.10
DM462Q01C	0.01	1.47	0.79	1.59	1.35
CM442Q02S	0.01	1.50	0.90	1.41	1.10
CM803Q01S	0.02	1.68	0.87	1.79	2.16
CM411Q02S	0.007	1.38	0.74	1.41	1.10
CM411Q01S	0.01	1.41	0.72	1.24	0.54
CM155Q04S	0	1.30	0.72	0.41	0.54
DM155Q03C	0.02	1.52	0.85	2.28	1.97
CM155Q01S	-0.007	1.45	0.53	-0.19	0.04
DM155Q02C	0.007	1.34	0.65	0.41	0.65
CM474Q01S	-0.005	1.29	0.65	0.28	0.04
CM033Q01S	0.005	1.35	0.62	-0.44	-0.35

APPENDIX D

Item difficulty parameters for Booklet 7 (for mother's education level)

Item	Primary School Dropout	Primary School	Middle School	High School	Undergraduate and Above
CM034Q01S	0.02	0.27	0.64	1.46	0.81
DM462Q01C	0.02	0.24	0.77	1.21	0.81
CM803Q01S	0.04	0.33	0.97	1.46	2.35
CM411Q02S	0.02	0.29	0.94	0.79	0.54
CM411Q01S	0.03	0.25	0.52	0.89	0.46
CM155Q04S	0.02	0.22	0.26	-0.36	-0.41
DM155Q03C	0.07	0.38	0.94	2.28	1.98
CM155Q01S	-0.02	0.18	0.19	0.15	-1.43
DM155Q02C	0.01	0.26	0.53	0.69	0.30
CM474Q01S	-0.01	0.26	0.31	-0.19	-0.09
CM033Q01S	-0.003	0.04	0.01	-0.82	-1.11
CM447Q01S	0.008	0.17	0.05	-0.82	-0.41
CM273Q01S	0.01	0.25	0.43	0.99	0.06
CM408Q01S	-0.008	0.26	0.47	0.41	0.30
CM420Q01S	0.004	0.26	0.36	-0.10	0.38
CM446Q01S	-0.03	0.10	0.22	-1.01	-1.43
DM446Q02C	0.13	0.82	2.02	2.51	3.27
CM559Q01S	-0.003	0.23	0.10	-0.36	-0.49
DM828Q02C	-0.01	0.23	0.15	-0.02	-0.58
CM828Q03S	0.04	0.36	0.87	1.33	1.10
CM464Q01S	0.06	0.25	0.89	1.74	1.10
CM800Q01S	-0.09	-0.27	-0.10	-3.12	-2.32