Eurasian Journal of Educational Research 68 (2017) 221-238

Eurasian Journal of Educational Research
www.ejer.com.tr

Cross-group Equivalence of Interest and Motivation Items in PISA 2012 Turkey Sample

Elif Ozlem ARDIC[1] Selahattin GELBAL[2]

ARTICLE INFO

Article History:
Received: 11 March 2015
Received in revised form: 02 December 2016
Accepted: 22 January 2017
DOI: http://dx.doi.org/10.14689/ejer.2017.68.12

Keywords
PISA
measurement invariance
multiple-group confirmatory factor analysis,
differential item functioning.

ABSTRACT

Purpose: The aim of this study was to examine measurement invariance of the interest and motivation related items contained in the PISA 2012 student survey with regard to gender school type and statistical regions and to identify the items that show differential item functioning (DIF) across groups. **Research Methods:** Multiple-group confirmatory factor analysis was conducted to examine measurement invariance. When the invariance with regard to gender was being investigated, potential item biases were examined, as the criteria used in the model fit evaluation were not met. Mantel-Haenszel, poly-SIBTEST, and item response theory likelihood ratio (IRT-LR) techniques were employed to identify which items displayed DIF. **Findings:** Results of the invariance test conducted based on the school type and statistical regions demonstrated that the models satisfied all invariance conditions. Failure to achieve measurement invariance according to gender indicates that at least one of the items in the scale displayed DIF. When the results of DIF according to gender were examined, MH identified DIF in six items at A level, poly-SIBTEST identified DIF in one item at A level, two items at B level, and three items at C level, IRT-LR identified DIF in two items at C level. **Implications for Research and Practice**: Further studies could determine which techniques would be more suitable for which situations by conducting simulation studies along with real data, and explore the possible reasons why the items display DIF.

[1] Hacettepe University, TURKEY
[2] Hacettepe University, TURKEY
Corresponding Author: Elif Özlem ARDIÇ, Hacettepe University, Faculty of Education, ardic@hacettepe.edu.tr

Introduction

The Programme for International Student Assessment (PISA) emphasizes factors that can affect student performance in addition to the school success of the students. To this end, student surveys have been conducted and student profiles have been formed to interpret the reasons behind PISA results. Conducting the application at certain intervals enables countries to compensate for their deficiencies and to monitor to what degree they realized their basic goals regarding education (Ministry of National Education [MNE], 2010). PISA is one of the applications through which participant countries change their education systems based on the results obtained. The meaningfulness of these results depends on the equivalence of the measurement tool forms across different groups. Employment of the measurement tools in which the demographic properties of the individuals are not disregarded, and comparative interpretation of the results obtained via these measurement tools might yield inaccurate results (Reise, Widaman & Pugh, 1993). This lowers the validity of the measurement tool and causes both affective and cognitive characteristics of the students to be inaccurately determined (Atalay Kabasakal & Kelecioğlu, 2012).

Validity of group comparisons depends on whether the relevant measurements possess an acceptable level of psychometric characteristics (Önen, 2007). However, in the classical test theory (CTT), test and item statistics calculated within the scope of validity and reliability studies reflect the properties of the study group (Crocker & Algina, 1986; Linden & Hambleton, 1997). In cases that individuals upon whom the measurement tool was applied differ with regard to factors such as geography, language, ethnicity, race, gender, etc., the same characteristics may not be measured similarly (Prelow, Tein, Roosa & Wood, 2000). This limitation of CTT constitutes the basis for the measurement invariance (Vandenberg & Lance, 2000).

Measurement Invariance

Flowers, Raju and Oshima (2002) defined measurement invariance as "the state that the individuals who are the members of different groups but have the same scores regarding a specific latent structure earn the same observed scores at the levels of items and sub-scales." According to this definition, measurement invariance can be expressed as the probability that an individual with a certain observed score does not depend on the group of the individual. Measurement invariance consists of steps, and at each step, an ever-increasing number of inter-group equality limitations are imposed with regard to the relevant parameters (Önen, 2009). The four steps proposed by Meredith (1993) and the hypothesis created for each step are as follows:

1) *Configural Invariance:* In this step, across-group equivalence limitation is imposed on the model, the theory of which was established (Wu, Li & Zumbo, 2007). Evidence of configural invariance means that the measurement tool represents the same psychological structure across groups (Vandenberg & Lance, 2000).

2) *Metric Invariance:* In addition to factor structure, factor loadings should also be equivalent in the sub-groups (Cheung & Rensvold, 2002). Ensuring metric invariance

shows similar/the same meaningfulness levels of the items for all groups (Johnson, 1998).

3) *Scalar Invariance:* In addition to factor structure and factor loadings, regression constants should also be equivalent across sub-groups in order to ensure scalar invariance (Cheung & Rensvold, 2002). It is necessary to achieve scalar invariance to compare the latent structure means across groups (Meredith, 1993).

4) *Strict Invariance:* In this step, it is hypothesized that error variances are equivalent across comparison groups.

In order for comparisons to be meaningful across groups, it is necessary to ensure measurement invariance (Van de Vijver & Tanzer, 2004). It would not be possible to figure out if the difference observed in the comparisons that are conducted without satisfying this requirement results from a real condition, or because the construct being measured differs across groups (Somer, Korkmaz, Dural & Can, 2009). Therefore, comparison results might be controversial.

Differential Item Functioning

A critical issue being discussed within the scope of measurement invariance across groups investigation is "bias" (Önen, 2009). Bias is defined as the systematic error against a group on the measurement results, and it affects the validity of the test scores (Angoff, 1993; Camilli, 2006). It is possible to investigate bias at item level via differential item functioning (DIF).

DIF, which is the first step of determining item bias, means the probability that responders with the same skill level in different groups answer the items in a test correctly differs (Holland & Wainer, 1993). DIF can also be described as the presence of the dimensions other than the construct aimed to be measured via the measurement tool (Roussos & Stout, 1996). The presence of DIF might misguide the researchers concerning the differences across groups and causes wrong decisions to be made about the individuals (Gök, Atalay Kabasakal & Kelecioğlu, 2014). In order to overcome this problem, studies on DIF are being carried out.

Based on the explanations provided so far, it is considered that the obtained data about the assessment of Turkey from the PISA application, in which member countries of the Organization for Economic Cooperation and Development (OECD) initially participated, has become one of the most significant research initiatives carried out worldwide today (MNE, 2010). As such, it is important to investigate measurement invariance and determining the items that show DIF across groups.

Purpose of Study

The aim of this study was to examine measurement invariance of the interest and motivation-related items contained in the PISA 2012 student survey with regard to gender, school types, and statistical regions, and to identify the items that show DIF across groups.

Method

Research Design

In this study, measurement invariance of eight items related to interest and motivation for mathematics involved in the PISA 2012 application was analyzed with regard to gender, school type, and statistical regions, and it was determined whether the items demonstrated DIF across genders. In terms of this, the study is descriptive and aims to determine an existing situation concerning psychometric characteristics of the measurements obtained from interest and motivation sub-scales.

Research Sample

Turkey participated in the PISA 2012 application with 4,848 students who represented approximately 1,266,638 students at the age of 15 (MNE, 2013). Following the investigation of the data set in terms of missing data and outliers, this study was carried out with 3,124 students (1,553 girls and 1,571 boys) in the Turkey sample. Table 1 presents the distribution of the students in the study group according to their school types and statistical regions.

Table 1

Distribution of the Students in Study Group According to School Types and Statistical Regions

	Primary School	General High School	Anatolian and Science High Schools	Technical and Vocational High Schools	Total
İstanbul Region	3	188	95	228	514
Western Marmara Region	2	10	21	83	116
Aegean Region	1	48	125	207	381
Eastern Marmara Region	1	57	62	196	316
Western Anatolia Region	1	99	66	168	334
Mediterranean Region	3	163	94	140	400
Central Anatolia Region	2	24	67	78	171
Western Black Sea Region	3	66	44	57	170
Eastern Black Sea Region	7	23	21	89	140
North Eastern Anatolia Region	6	23	44	40	113
Central Eastern Anatolia Region	11	81	21	41	154
South Eastern Anatolia Region	27	151	50	87	315
Total	67	933	710	1414	3124

Data Collection

In this study, data obtained from Turkey sample in the student survey, which was administered within the scope of PISA, organized by OECD in 2012, were used. The investigations were carried out through the answers provided to eight items related to interest and motivation in the student survey within the scope of the study. The data used in the study were obtained from the OECD PISA website (www.pisa.oecd.org).

Data Analysis

With a view to obtaining evidence regarding whether ST29Q01-ST29Q08 items contained in the PISA student survey mathematics teaching sub-dimension created the interest and motivation model, confirmatory factor analysis (CFA) was executed. The data set was examined prior to the analysis and the analysis run revealed that missing data rates regarding each variable varied between 0.78 and 1.25. These data were excluded from the analysis since the missing data rate was found to be less than 5% (Tabacknick & Fidell, 2007; Kline, 2011, p. 55).

Distribution characteristics of the relevant data set were examined in order to determine which parameter prediction method would be employed during the model testing process. To this end, z values regarding multivariate skewness (zs), kurtosis (zk), and $\chi 2$ value (zs=24.80, zk=34.982 and $\chi 2$=1842.793, $p<.05$) regarding multivariate skewness and kurtosis were calculated. Since the data set was not normally distributed and the sample size was large, the weighted least squares (WLS) method was used in parameter prediction (Kline, 2011, p. 180).

Multiple-group confirmatory factor analysis (MG-CFA) was conducted in order to examine measurement invariance. The analysis started with testing the least limited model and continued by increasing the number of limitations. With the aim of comparing the fit levels of a more limited and less limited model with the research data, the scaled difference chi-square test was applied (Bentler; 2006; Brown, 2006). $\Delta S\text{-}B\chi 2$ calculated based on the difference between the degrees of freedom of two models was not found to be statistically significant, and this was interpreted as evidence that invariance was achieved at that level (Vanderberg & Lance, 2000; Byrne & Watkins, 2003; Mark & Wan, 2005).

When the invariance with regard to gender was being investigated, potential item biases were examined since the criteria used in the model fit evaluation were not met. Mantel-Haenszel, poly-SIBTEST, and item response theory likelihood ratio (IRT-LR) techniques were employed to identify which items displayed DIF. Mantel-Haenszel analysis was conducted via JMETRIK, Poly-SIBTEST was conducted via SIBTEST, and IRT-LR analysis was conducted via IRTLRDIF software. The statistics taken into consideration to identify the items that demonstrated DIF were p for MH, βu for SIBTEST, and G2 for IRT-LR.

Results

Step I: Confirmatory Factor Analysis

In order to examine whether the factor structure of the basic model described regarding the factor structure of the 8-item Turkish form of the survey was valid within each group, the fit of the model was examined separately for the integrated data and for the data of each group. Table 2 summarizes the fit indexes calculated for each group at the end of the CFA analysis.

Table 2

Interest and Motivation Measurement Model Fit Index

GROUP	$\chi 2$	*df*	*RMSEA*	*GFI*	*AGFI*	*CFI*	*NNFI*
Whole Group	199.05	19	0.055	0.98	0.96	0.99	0.99
Female	195.83	19	0.077	0.96	0.92	0.85	0.78
Male	75.64	19	0.044	0.98	0.96	0.95	0.93
Primary School	32.35	19	0.103	0.83	0.67	0.97	0.95
General High Schools	71.89	19	0.055	0.97	0.94	0.99	0.99
Anatolian and Science High Schools	97.92	19	0.077	0.96	0.92	0.99	0.98
Tech. and Vocational High Schools	87.72	19	0.049	0.98	0.96	0.99	0.99
İstanbul Region	41.38	19	0.048	0.97	0.94	0.95	0.93
Western Marmara Region	31.06	19	0.074	0.93	0.87	0.91	0.87
Aegean Region	26.98	19	0.033	0.97	0.95	0.98	0.96
Eastern Marmara Region	33.67	19	0.050	0.96	0.93	0.99	0.99
Western Anatolia Region	41.66	19	0.060	0.96	0.92	0.99	0.99
Mediterranean Region	30.80	19	0.039	0.97	0.95	0.97	0.95
Central Anatolia Region	31.31	19	0.062	0.95	0.91	0.93	0.90
Western Black Sea Region	31.57	19	0.063	0.92	0.85	0.99	0.99
Eastern Black Sea Region	30.28	19	0.065	0.93	0.86	0.99	0.98
North Eastern Anatolia Region	19.69	19	0.018	0.97	0.94	1.00	0.99
Central Eastern Anatolia Region	33.38	19	0.070	0.93	0.88	0.99	0.98
South Eastern Anatolia Region	40.33	19	0.060	0.96	0.91	0.99	0.99

*p<.05

When the fit statistics related to the relevant model are examined, it is seen that the criteria used in the model fit assessment are within the acceptable boundaries (GFI=0.98, AGFI=0.96, CFI=0.99, NNFI=0.99, and RMSEA=0.055). Factor loading values and unique variances were found to vary between 0.75 and 0.88 and between 0.23 and 0.44, respectively. When the fit criteria regarding the model in the sub-groups were examined, the values with regard to NNFI for the girls; RMSEA, GFI, and AGFI for the primary schools; and NNFI for the Western Marmara Region did not meet the criteria required for the model fit. Therefore, it was decided to exclude these three groups that did not achieve the model data fit from the analysis.

Step II: Multiple-Group Confirmatory Factor Analysis

A four-step method was followed in order to examine the measurement invariance of the defined method across different groups. The findings obtained from the analysis carried out during measurement invariance investigation process were interpreted in line with the research questions. The results of the measurement invariance analysis conducted on the school types are as follows:

Configural Invariance: In this step, whether the factor structures of groups are equivalent was tested within the same model. The analysis results revealed that the fit indexes were within acceptable boundaries (CFI=0.96, GFI=0.98, NNFI=0.95, RMSEA=0.037, and S-Bχ2 =173.94 (df=73), and this indicated that configural invariance was ensured. This means that the groups that provided the answers had the same conceptual points of view.

Metric Invariance: The fit indexes obtained after imposing equivalent factor loadings limitation within the groups along with configural invariance steps limitation show that metric invariance model fits to the relevant data at a satisfactory level (CFI=0.96, GFI=0.98, NNFI=0.96, RMSEA=0.032, and S-Bχ2 =173.96 (df=85)). In order to provide evidence that metric invariance was ensured, the fit level of this model and the fit level of configural invariance model were compared via the scaled difference chi-square test. TS statistic calculated via the scaled difference test was found to be smaller than the Table χ2 value (χ2(12, .05)=21.03) for df=12, and this indicates that metric invariance was achieved. This signifies that meaning of the items is similar to the students at different schools.

Scalar Invariance: In addition to the limitations established in the first two steps, regression constants were also limited. Fit indexes calculated in order to analyze scalar invariance are CFI=0.96, GFI=0.97, NNFI=0.96, RMSEA=0.032, and S-Bχ2=182.66 (df=90). When the fit level of this model and scalar invariance model were compared, TS statistic (TS=9.022) was found to be smaller than the Table χ2 value (χ2(5, .05)=11.07), and this indicates that predicted item scores were obtained irrespective of the group membership. In other words, the items did not display bias.

Strict Invariance: Error variances are limited together with all previous parameter limitations. As a result of the MG-CFA analysis conducted in order to test strict invariance, fit indexes were found to be CFI=0.96, GFI=0.97, NNFI=0.96, RMSEA=0.032, and S-Bχ2 =182.66 (df=91). The TS statistic TS=0 (df=1) calculated was

found to be smaller than Table $\chi 2$ value ($\chi 2(1, .05)=3.841$), and this indicates that error variances do not differ depending on the school types.

Results of the measurement invariance analysis of the defined measurement model across statistical regions are as follows:
Configural Invariance: Results of MG-CFA analysis conducted to test configural invariance revealed that fit indexes calculated were within acceptable boundaries (CFI=0.92, GFI=0.94, NNFI=0.92, RMSEA=0.043, and S-B$\chi 2$=469.62 (df=313), and this indicates that configural invariance was ensured.

Metric Invariance: It can be inferred that after imposing equivalent factor loadings limitation within the groups, it was ensured that the model fits to the relevant data at a satisfactory level (CFI=0.95, GFI=0.94, NNFI=0.96, RMSEA=0.031, and S-B$\chi 2$=469.63 (df=373)). TS=0 and df=60 values were obtained following the scaled difference in the $\chi 2$ test. TS statistic calculated was found to be smaller than the table $\chi 2$ value (79.08), and this indicates that metric invariance was achieved.

Scalar Invariance: Fit indexes calculated in order to analyze scalar invariance are CFI=0.95, GFI=0.94, NNFI=0.96, RMSEA=0.030, and S-B$\chi 2$ =470.58 (df=378). When the fit level of the scalar invariance model and metric invariance model were compared, TS statistic was compared to the table $\chi 2$ value ($\chi 2(5, .05)=11.071$) for df=5, and TS statistic (TS=0.587) was found to be smaller than the Table $\chi 2$ value. This indicates that scalar invariance was ensured.

Strict Invariance: MG-CFA executed in order to analyze the invariance of error variances presented fit indexes as CFI=0.96, GFI=0.94, NNFI=0.96, RMSEA=0.030, and S-B$\chi 2$ =470.58 (df=379). The Ts=0 statistics calculated as a result of the scaled difference chi-square test for $\chi 2$ was smaller than the Table $\chi 2$ ($\chi 2$ (1, .05)=3.841) value, which shows that invariance of error variances had ensured.

Step III. Determining the Items Demonstrating DIF According to Gender

Failure to achieve measurement invariance according to gender indicates that at least one of the items in the scale displayed DIF. It is seen that different techniques employed in determining DIF yields different items with DIF. For this reason, it is recommended that numerous methods be used in the DIF analysis (Hambleton, 2006). Accordingly, MH, poly-SIBTEST, and IRT-LR techniques were employed to determine if items showed invariance across genders, and results were compared. Regarding the items that displayed DIF across genders, the results of the MH technique, the poly-SIBTEST technique, and the IRT-LR technique are found in Tables 3, 4, and 5, respectively.

Table 3

MH Analysis Results of Interest and Motivation Items According to Gender Variable

Item	$\chi 2$	*p*	Δ-*MH*	*DMF Level*
ST29Q01	17.43	0.00	-0.10	A
ST29Q02	29.12	0.00	0.10	A
ST29Q03	10.74	0.00	-0.06	A
ST29Q04	0.79	0.37	-0.02	
ST29Q05	16.27	0.00	0.08	A
ST29Q06	11.07	0.00	-0.06	A
ST29Q07	5.03	0.02	0.05	A
ST29Q08	0.99	0.77	0.01	

Reference group: males; Focus group: females

MH results indicate that a negligible level (A level) of DIF was presented in six items.

Table 4

Poly-SIBTEST Analysis Results of Interest and Motivation Items According to Gender Invariable

Item	βu	Standard Error	p	DMF Level	Advantageous Group
ST29Q01	-0.108	0.025	0.000	C	K
ST29Q02	0.126	0.021	0.000	C	E
ST29Q03	-0.085	0.023	0.000	B	K
ST29Q04	-0.013	0.022	0.548		
ST29Q05	0.100	0.024	0.000	C	E
ST29Q06	-0.069	0.022	0.002	B	K
ST29Q07	0.054	0.023	0.018	A	E
ST29Q08	0.008	0.025	0.765		

Reference group: male students; Focus group: female students

Table 4 presents that the beta value is significant in six items out of eight. When these six items are examined, it is seen that one of them displayed DIF at A level, two of them at B level, and three at C level. The item that displayed DIF at A level favored boys, whereas the items that displayed DIF at B level favored girls. Of the

items that displayed DIF at C level, ST29Q01 showed DIF in favor of girls, and ST29Q02 and ST29Q05 showed DIF in favor of boys.

Table 5

IRT-LR Analysis Results of Interest- and Motivation-Related Items According to Gender Variable

		parameter			
Items	G2	*A*	*b*	*c*	*DIF Level*
ST29Q01	15.9		K		B
ST29Q02	0.0				
ST29Q03	15.7		E		B
ST29Q04	0.0				
ST29Q05	0.0				
ST29Q06	0.0				
ST29Q07	3.1				
ST29Q08	0.0				

Reference group: male students; Focus group: female students

When interest and motivation scale items were analyzed via the IRT-LR technique with respect to gender variable, B level of DIF was observed in two items. Item ST29Q01 showed DIF favored female students, while item ST29Q03 showed DIF favored male students.

The distribution of the items that displayed DIF in each of the three methods according to the levels at the end of the analysis run via MH, Poly-SIBTEST, and IRT-LR techniques are presented in Table 6.

Table 6

Distribution of the Items that Displayed DIF According to Gender

MH			Poly-SIBTEST			IRT-LR		
A	B	C	A	B	C	A	B	C
ST29Q01			ST29Q07	ST29Q03	ST29Q01		ST29Q01	
ST29Q02				ST29Q06	ST29Q02		ST29Q03	
ST29Q03					ST29Q05			
ST29Q05								
ST29Q06								
ST29Q07								

When Table 6 is examined, it is seen that MH identified DIF in six items, Poly-SIBTEST identified DIF in six items, and IRT-LR identified DIF in two items. In these methods, two items (ST29Q01 and ST29Q03) showed DIF. However, two items that displayed DIF at A level in the MH method displayed DIF at B level in the IRT-LR

method. Furthermore, according to the Poly-SIBTEST method, item ST29Q01 showed DIF at C level, and item ST29Q03 showed DIF at B level.

There are four items that did not display DIF via the IRT-LR method, but showed DIF via MH and Poly-SIBTEST methods. These are items ST29Q02, ST29Q05, ST29Q06, and ST29Q07, which display DIF according to the MH method at A level, and according to the Poly-SIBTEST method at A (ST29Q07), B (ST29Q06), and C (ST29Q02 and ST29Q05) levels.

As a result, it is clear that two items (ST29Q01 and ST29Q03) displayed DIF via all these three methods, however, their levels are different. The MH and Poly-SIBTEST methods showed fit in identifying DIF; however, their levels were found to be different. Different from the two methods, the IRT-LR method, however, shows that only 25% of the items displayed DIF.

Discussion and Conclusion

Findings of the study indicated that the model described with regard to the interest and motivation-related items contained in the PISA 2012 student survey Turkish form sufficiently fits all sub-group data, except for female students, primary schools, and Western Marmara groups. After the groups that did not fit the model were excluded from the analysis, equality of content was ensured among the sub-groups (Önen, 2009).

Results of the invariance test conducted based on the school type demonstrated that the model satisfied all the invariance conditions. This signifies that the measurements obtained from all interest and motivation-related items could be generalized among the school groups, and provide reliable and valid measurements in determining the interests and motivations of the students. Nevertheless, ensuring a complete measurement invariance among all groups is not always possible (Steenkamp & Baumgartner, 1998). Likewise, Uyar and Doğan (2014), found that the model they described for learning strategies met the configural and metric invariance conditions in the sub-groups.

The analysis indicated that comparison of the described model according to statistical regional was significant. Accordingly, it could be said that the difference observed in the comparisons in the regional groups resulted from the real situation. This finding is in parallel with the study by Uyar and Doğan (2014) that investigated the differences of the variable affecting learning strategies across regions. Similarly, Wu et al. (2007) specified that TIMSS 1999 mathematics tests ensured strict invariance in the same cultures. Numerous studies investigating the sub-group invariance of different models that were described regarding the international large-scaled exams showed that all invariance steps were not ensured (Wu et al., 2007; Akyıldız, 2009; Uzun & Öğretmen, 2010).

The spread of the large-scaled exams paved the way for different test forms to be administered to individuals at the same level, and for the same test forms to be

administered in groups with different characteristics (Atalay Kabasakal, 2014). Within this scope, a point that must be considered in administering national and international tests is the impact of the membership of different demographic groups on the measurement results. The national test applications performed at the national level indicate that the reasons for DIF include variables such as gender and school type (Bakan Kalaycıoğlu & Kelecioğlu, 2011; Gök, Kelecioğlu & Doğan, 2010). Le (2009) maintained that inclusion of items that displayed DIF in the internationally large-scaled exams such as PISA is inevitable. Similarly, the results of the present study also indicated that gender difference affected that the items displayed DIF. Similarly, the studies by Le (2009), Atalay Kabasakal and Kelecioğlu (2012), Akın Arıkan (2015), Başokçu and Öğretmen (2013) found that the items in the test applications displayed DIF across genders. When gender-related DIF is examined, it is seen that the characteristics of items such as format, scope, and cognitive complexity level are seen among the possible reasons for DIF (Bakan Kalaycıgolu & Berberoglu, 2010; Zumbo & Gelin, 2005, Mendes-Barnett & Ercikan, 2006). Contrary to the results of this study, Başusta & Gelbal (2015), however, presented that the science and technology items in the PISA 2006 student survey could provide valid and reliable measurements across genders.

Although DIF identification techniques provide similar results at certain levels in a general sense, since they use different equalization criteria, algorithms and breakpoints in categorizations, they are not in a complete fit (Bakan Kalaycıoglu & Berberoglu, 2010). In accordance with these findings, it was observed that similarities between the number of items displaying DIF and amount of DIF was low according to the techniques used. Similarly, studies by Gök, Kelecioğlu and Doğan (2010), Çıkrıkçı Demirtaşlı and Uluştaş (2015) found a difference between the techniques in terms of the number of items that showed DIF. The analysis results showed that the number of the items with DIF was found to be high via MH and Poly-SIBTEST techniques. This may have resulted because these techniques are more sensitive compared to the IRT-LR technique. In addition, these techniques require smaller samples than the techniques based on the Item Response Theory, which could be seen as an advantage of these techniques (Penfield & Camilli, 2007). The studies have demonstrated that the reasons for the differences among the techniques include factors such as the various difficulties and discrimination of the items, different sample sizes, different group means, and skills (Hidalgo & Pina, 2004; Narayanan & Swaminathan, 1996; Fidalgo, Mellenbergh & Muñiz, 2000).

Within the scope of this study, interest and motivation-related items contained in the PISA 2012 application mathematics teaching section were examined. Further studies might examine the measurement invariance of the survey items administered within the scope of international studies such as PISA, PIRLS, and TIMSS for groups with differing cultures and religions. This study employed MH, poly-SIBTEST, and IRT-LR techniques in order to identify the items that displayed DIF. Future studies could determine which techniques would be more suitable for which situations by conducting simulation studies along with real data, and exploring the possible reasons why the items display DIF.

References

Akın Arıkan, Ç. (2015). Değişen madde fonksiyonu belirlemede mtk-olabilirlik oranı, ordinal lojistik regresyon ve poly-sibtest yöntemlerinin karşılaştırılması [Comparison of irt likelihood ratio test, poly-sibtest and logistic regression difdetection procedures]. *Uluslararası Eğitim Araştırmaları Dergisi*, 6(1), 1-16.

Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkeler arası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 6(1), 18-47.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum.

Atalay Kabasakal, K. & Kelecioğlu, H. (2012). PISA 2006 Öğrenci anketinde yer alan maddelerin değişen madde fonksiyonu açısından incelenmesi [Evaluation of attitude items in PISA 2006 student questionnaire in terms of differential item functioning]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 45(2), 77-96.

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi [The effect of differential item functioning on test equating*]. Unpublished doctoral thesis, Hacettepe University, Ankara.

Bakan Kalaycıoğlu, D. & Berberoğlu, G. (2010). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment*, 20, 1-12.

Bakan Kalaycıoğlu, D. & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item bias analysis of the university entrance examination]. *Eğitim ve Bilim*, 36, 3-13.

Başokçu, T. & Öğretmen, T. (2013). Öğretmen öz yeterlilik ölçeğinde değişen madde fonksiyonlarının ağırlıklandırılmış cevap modeli ile belirlenmesi [Determine the differential item functioning in teacher self efficacy by graded response model]. *Ege Eğitim Dergisi*, 14(2), 63-78.

Başusta, N. B & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement invariance at groups' comparisons: a study on PISA student questionnaire]. *Hacetepe Üniversitesi Eğitim Fakültesi Dergisi*, 30(4), 80-90.

Bentler, P. M. (2006). EQS 6 Structural equations program manual. Encine, CA: Multivariate Software, Inc.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Bryne, B. M. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*. 34(2), 155-175.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport: American Council on Education&Praeger Publishers.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255.

Çıkrıkçı Demirtaşlı, N. & Uluştaş, S. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research*, 58, 41-60.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory. Orlando*: Harcourt Brace JovanovichInc.

Fidalgo, A. M., Mellenbergh, G. J. & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of mantel-haenszel procedures. *Methods of Psychological Research*, 5(3), 43-53.

Flowers, C.P., Raju, N. S. & Oshima, T.C. (2002). A comparison of measurement equivalence methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517–529.

Gök, B., Atalay Kabasakal, K. & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA 2009 student questionnaire in terms of differential item functioning based on culture]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(1), 72-87.

Gök, B., Kelecioğlu, H. & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenzsel ve lojistik regresyon tekniklerinin karşılaştırılması [The comparison of mantel-haenszel and logistic regression techniques in determining the differential item functioning]. *Eğitim ve Bilim*, 35(156), 3-16.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44, 182-188.

Higaldo, M. D. & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect-size: a comparison between LR and MH procedures. *Educational and Psychological Measurement,* 64(6), 903–915.

Holland, P. W. & Wainer, H. (1993).Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, T. P. (1998). Approaches to equivalence in crosscultural and cross-national survey research. *ZUMA-Nachrichten Spezial*, 1-40.

Kline, R. B. (2011). *Principles and practice of structural equation modelling* (3rd Edition). New York: Guildford Publication, Inc.

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

Linden, V. D. & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-VerlagInc.

Mark, B. A. & Wan, T.T.H (2005). Testing measurement equivalence in a patient satisfaction instrument. *Western Journal of Nursing Research*,27 (6), 772-787.

Mendes-Barnett, S. & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessment susing a confirmatory multidimensional model approach. *Applied Measurement in Education*, 19, 289-304.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.

MNE (2010). *PISA 2009 Uluslararası Öğrenci Değerlendirme Programı Ulusal Ön Raporu*. MEB, Ankara.

MNE (2013). PISA 2012 Uluslararası Öğrenci Değerlendirme Programı Ulusal Ön Raporu. MEB, Ankara.

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

Önen, E. (2007). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin incelenmesi: epistemolojik inançlar envanteri üzerine bir çalışma [Examination of measurement invariance at groups' comparisions: a study on epistemological beliefs inventory]. *Ege Eğitim Dergisi*, 2(8), 87–110.

Önen, E. (2009). *Ölçme değişmezliğinin yapısal eşitlik modelleme teknikleri ile incelenmesi* [Examination of measurement invariance with structural equation modelling techniques]. Unpublished doctoral thesis, Ankara University, Ankara.

Organization for Economic Cooperation and Development Programme for International Student Assessment Web Site. Retrieved November 20, 2015, from http://www.pisa.oecd.org

Penfield, R. D. & Camilli, G. (2007). Dierential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics Psychometrics* (26, pp. 125–167). Amsterdam: Elsevier.

Prelow, H. M., Tein, J.Y., Roosa, M. W. & Wood, J. (2000). Do coping styles differ across sociocultural groups? The role of measurement equivalence in making this judgment. *American Journal of Community Psychology*, 28 (2), 225-244.

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

Roussos, L.L. & Stout, W. F. (1996). Simulation studies of the effects of small sample

size and studied item parameters on sibtest and mantel-haenszel type I error performance. *Journal of Educational Measurement*, 33(2), 215–230.

Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Detection of measurement equivalence by structural equation modeling and item response theory. *Turkish Journal of Psychology*, 24(64).

Steenkamp, E. M & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research*, 25(1), 78-90.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th Edition). Boston MA: Allyn& Bacon.

Uyar, Ş. & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample]. *Uluslararası Türk Eğitim Bilimleri Dergisi,* 2(3), 30-43.

Uzun, B. & Öğretmen, T. (2010). Fen başarısı ile ilgili bazı değişkenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey sample]. *Eğitim ve Bilim*, 35(155), 26-35.

Van de Vijver, F. J. R. & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, 54, 119-135.

Vandenberg, R.J. & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.

Wu, D. A., Li, Z. & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assesment, Research & Evaluation*, 12(3),1-26.

Zumbo, B. D. & Gelin, M. N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1-23.

PISA 2012 Türkiye Örneklemi İlgi ve Motivasyon Maddelerinin Gruplar Arası Karşılaştırmalarda Eşdeğerliğinin İncelenmesi

Atıf

Ardic, E.O. & Gelbal, S. (2017). Cross-group Equivalence of Interest and Motivation Items in PISA 2012 Turkey Sample. *Eurasian Journal of Educational Research*, 68 221-238, http://dx.doi.org/10.14689/ejer.2017.68.12

Özet

Problem Durumu: Grup karşılaştırmaların geçerliği, ilgili ölçümlerin kabul edilebilir düzeyde psikometrik niteliklere sahip olmasına bağlıdır. Ancak klasik test kuramında, geçerlik ve güvenirlik çalışmaları kapsamında hesaplanan test ve madde istatistikleri, araştırma grubunun özelliklerini yansıtmaktadır. Büyük ölçekli sınavların kullanımının yaygınlaşması ise aynı düzeydeki bireylere farklı test formlarının uygulanmasına ve aynı test formlarının farklı özelliklere sahip gruplarda uygulanmasına yol açmıştır. Bu bağlamda ulusal ve uluslararası test uygulamalarında dikkat edilmesi gereken bir durum, farklı demografik gruplara ait olmanın ölçme sonuçları üzerindeki etkisidir. Bireylerin demografik özelliklerinin etkisinin arındırılmadığı ölçme araçlarının kullanılması ve bunlardan elde edilen sonuçların karşılaştırılmalı olarak yorumlanması ise ölçme aracının geçerliğini düşürerek, bireyler hakkında yanlış kararlar alınmasına neden olacaktır. Bu nedenle ölçme sonuçlarına dayalı olarak verilecek kararların isabetliliği açısından ölçme değişmezliğinin sağlanması ve maddelerin olası yanlılık şüphesine karşı sınanması gerekmektedir. Bu koşullar sağlanmadan yapılan karşılaştırmalarda görülen farklılığın gerçek durumdan mı yoksa ölçülen yapının gruplarda farklılık göstermesinden mi kaynaklandığı bilinemeyecektir. Dolayısıyla yapılan karşılaştırma sonuçları tartışmalı olabilecektir.

Araştırmanın Amacı: Bu çalışmanın amacı; PISA 2012 öğrenci anketinde yer alan ilgi ve motivasyonla ilgili maddelerin cinsiyet, okul türü ve istatistikî bölgelere göre ölçme değişmezliğini incelemek ve gruplar arası DMF gösteren maddeleri tespit etmektir.

Araştırmanın Yöntemi: PISA 2012 uygulamasında Türkiye, 15 yaş grubu yaklaşık sayısı 1.266.638 öğrenciyi temsilen 4848 öğrenci ile yer almıştır. Veri setinin kayıp ve aykırı değerler açısından incelenmesi sonrasında bu araştırma, Türkiye örneklemindeki 3124 öğrenci (1553 kız ve 1571 erkek) ile yürütülmüştür. PISA öğrenci anketi matematik öğretimi alt boyutunda yer alan ST29Q01-ST29Q08 maddelerinin ilgi ve motivasyon modelini oluşturup oluşturmadığına ilişkin kanıtlar elde etmek üzere, doğrulayıcı faktör analizi uygulanmıştır. Ölçeğin 8 maddelik Türkçe formunun faktör yapısına ilişkin tanımlanan temel modelin faktör yapısının her bir grup içinde geçerli olup olmadığını incelemek için model uyumu birleştirilmiş veri ve her bir grup verisi için ayrı ayrı değerlendirilmiştir. Model test etme sürecinde, hangi parametre kestirim yönteminin kullanılacağını belirlemek için ilgili veri setinin dağılım özellikleri incelenmiştir. Veri seti çok değişkenli normal dağılım sergilemediği ve örneklem sayısı büyük olduğu için parametre kestiriminde

ağırlıklandırılmış en küçük kareler yöntemi kullanılmıştır. Ölçme değişmezliğini incelemek üzere çoklu grup doğrulayıcı faktör analizi uygulanmıştır. Değişmezlik testleri dört aşamada yürütülmüştür. Daha fazla sınırlama konulan bir model ile daha az sınırlama konulan bir modelin araştırma verisine uyum düzeylerini karşılaştırmak üzere $\chi 2$'ler için ölçeklendirilmiş fark testi uygulanmıştır. Cinsiyete göre ölçme değişmezliğinin incelenmesi sürecinde, model uyumunun değerlendirilmesinde kullanılan ölçütler karşılanmadığı için olası madde yanlılıkları incelenmiştir. DMF gösteren maddelerin belirlenmesi amacıyla Mantel-Haenszel, poly-SIBTEST ve MTK-OO teknikleri kullanılmıştır.

Araştırmanın Bulguları: Ölçeğin 8 maddelik Türkçe formunun faktör yapısına ilişkin tanımlanan temel modelin kız öğrenci, ilköğretim ve Batı Marmara grupları dışındaki tüm alt grup verilerine yeterli düzeyde uyum sergilediğini göstermiştir. Modele uyumunu sağlamayan gruplar, analiz dışında bırakılmıştır. Okul türü ve istatistiki bölgelere dayalı olarak yapılan değişmezlik testi sonuçları, modellerin tüm değişmezlik koşullarını yerine getirdiğini göstermiştir. Cinsiyete göre ölçme değişmezliğinin sağlanmaması, ölçekte yer alan maddelerden en az bir tanesinin cinsiyete göre DMF sergilediğine işaret etmektedir. Bu bağlamda, cinsiyete göre DMF sonuçları incelendiğinde MH tekniğine göre 6 maddede A düzeyinde; poly-SIBTEST tekniğine göre 1 maddede A, 2 maddede B ve 3 maddede C düzeyinde; MTK-OO tekniğine göre 2 maddede C düzeyinde DMF'ye rastlanmıştır.

Araştırmanın Sonuç ve Önerileri: Yapılan analizler tanımlanan modelin, okul türü ve istatistiki bölgelere göre karşılaştırılmasının anlamlı olduğunu ortaya koymuştur. Bu durum, ilgi ve motivasyonla ilgili tüm maddelerden elde edilen ölçümlerin okul grupları ve istatistiki bölgeler arasında genellenebileceğine, öğrencilerin ilgi ve motivasyonlarını belirlemede geçerli ve güvenilir ölçümler sağlayabileceğine işaret etmektedir. Bu doğrultuda okul ve bölge grupları arasında yapılan karşılaştırmalarda görülen farklılığın gerçek durumdan kaynaklandığı söylenebilir. Yapılan ulusal düzeydeki test uygulamaları, DMF'nin nedenleri arasında cinsiyet ve okul türü gibi değişkenleri göstermektedir. Nitekim bu çalışmanın sonuçları da cinsiyet farklılıklarının maddelerin DMF göstermesinde etkili olduğunu göstermiştir. DMF belirleme teknikleri genel olarak belli ölçüde benzer sonuçlar verse de, farklı eşitleme kriterleri ile farklı algoritmalar ve kategorilendirmelerde farklı kesme noktaları kullandıkları için tam bir uyum içinde değildir Çalışmadan elde edilen sonuçlar incelendiğinde, kullanılan tekniklere göre DMF gösteren madde sayıları ve DMF miktarları arasındaki benzerliğin düşük düzeyde olduğunu gözlenmiştir. Bu çalışma kapsamında PISA 2012 uygulaması matematik öğretimi bölümünde yer alan ilgi ve motivasyonla ilgili maddeler incelenmiştir. Gelecek çalışmalar, farklı dil ve kültür grupları üzerinde ölçme değişmezliği çalışmaları yapabilir. DMF belirlemede gerçek veri ile birlikte simülasyon çalışmaları yaparak, hangi tekniğin hangi durumlar için daha uygun olduğunu belirleyebilir ve DMF gösteren maddelerin olası nedenlerini araştırabilir.

Anahtar Kelimeler: PISA, ölçme değişmezliği, çoklu grup doğrulayıcı faktör analizi, değişen madde fonksiyonu.