International Journal of Education & Literacy Studies
ISSN: 2202-9478
www.ijels.aiac.org.au

Cross-cultural Measurement Invariance of the Items in the Science Literacy Test in the Programme for International Student Assessment (PISA-2015)*

Betül Alatlı*

Department of Educational Science, Faculty of Education, Tokat Gaziosmanpasa University, Turkey

Corresponding author: Betül Alatlı, E-mail: betul.alatli@gop.edu.tr

ARTICLE INFO

Article history
Received: January 18, 2020
Accepted: April 07, 2020
Published: April 30, 2020
Volume: 8 Issue: 2

Conflicts of interest: None
Funding: None

*A part of this study was presented as an oral presentation at 6th International Congress on Measurement and Evaluation in Education and Psychology (September, 5-8, 2018 Prizren, KOSOVO).

ABSTRACT

This study aimed to investigate cross-cultural measurement invariance of the PISA (Programme for International Student Assessment, 2015) science literacy test and items and to carry out a bias study on the items which violate measurement invariance. The study used a descriptive review model. The sample of the study consisted of 2224 students taking the S12 test booklet from Australia, France, Singapore, and Turkey. Measurement invariance analyses for the test were done using Multi-Group Confirmatory Factor Analysis (MGCFA). Differential Item Functioning (DIF), in other words, measurement invariance of the test items, was analyzed using the item response theory log-likelihood ratio (IRTLR), Hierarchical Generalized Linear Model (HGLM), and the Simultaneous Item Bias Test (SIBTEST) methods.According to the findings, the test was determined to exhibit structural invariance across cultures. The highest number of items showing DIF was observed in the comparisons of Australia-Singapore and Australia-France with 35%. The number of items showing DIF, with 24%, determined in bilateral comparisons which included Turkey, the only country taking the translated form among other countries, did not show a significant difference compared to the other comparisons. While the lowest number of items showing DIF was obtained from Singapore-France samples with 12%, the rate of items indicating DIF in the France-Turkey samples was 18%. On the other hand, 35% of the items showed cross cultural measurement invariance. An item bias study was carried out based on expert opinions on items identified and released as showing DIF in the comparisons of Turkey with Australia and Singapore.According to the findings, translation-bound differentiation of the items, familiarity of a culture group with the contents of the items, polysemy in the expressions or words used in the items, the format, or the stylistic characteristics of the items were determined to be the cause of the bias in the skills measured with the items.

Key words: Differential Item Functioning, Item Bias, Measurement Invariance, Pisa, Science Literacy

INTRODUCTION

Education is defined as a permanent and multi-faceted change process aiming to provide individuals with prosperity and happiness (Demirtaşlı, 2014). Standard measurement and evaluation systems are needed to determine the level of change that education aims to achieve. Evaluation is the quality control system of the education process. For this reason, countries can determine the deficiencies in their education systems with standard measurement and evaluation outputs at national and international dimensions and obtain important feedback such as the level of behavioral change expected to be changed in students. This kind of feedback is considered very important in terms of guiding the educational policies of the country. In this sense, an educational survey carried out at the international level, which provides important outputs about the education systems of countries, is quite effective. One of these surveys is the PISA application, which is organized by the Organization for Economic Co-operation and Development (OECD). It measures students' science and mathematics literacy and reading skills as well as making student, teacher, and school-level measurements. With PISA, countries can make comparisons on an international scale, identify the lacking aspects of the current system, and guide their educational policies. PISA has been shown to have a significant impact on the educational policies of countries (Ercikan, Roth & Asil, 2015; Niemann, Martens & Teltemann, 2017; Sjøberg, 2015).

PISA is administered every three years. PISA 2015 focused on science literacy. PISA 2015 application involved 72 countries. For some countries, tests are adapted to multiple languages and cultures. Therefore, there were 82

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.ijels.v.8n.2p.16

country-language combinations in PISA 2015. For example, Estonian and Russian language forms of the tests are developed for Estonia. For such implementations, testing the cultural and linguistic measurement invariance, which is seen as the most important element of the culture, is considered to be very important. To make accurate and fair inferences about the results of international applications such as PISA, the measurement invariance must be met (Gierl, 2000; Vandenberg & Lance, 2000). Therefore, this study demonstrates the importance of investigating the measurement invariance across different cultures taking different and similar language forms of the science literacy test as the focus of PISA 2015 was on science literacy.

The necessity of examining the intercultural or lingual measurement invariance of measurement tools employed in educational research conducted at an international level is clearly shown in both Test Adaptation Guidelines (International Test Commission [ITC], 2005) and Measurement Standards in Education and Psychology (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999). Measurement invariance is defined as obtaining the same observed score at the item and subscale level when individuals in different groups have the same score in terms of a certain implicit structure (AERA, APA & NCME, 1999). Measurement invariance can be achieved by keeping the relationships between observed and latent variables the same for different groups. When the literature is examined, the most commonly used and recommended approaches for examining measurement invariance are Differential Test Functioning (DTF) and Differential Item Functioning (DIF) based on Item Response Theory (IRT), and Structural Equation Modeling (SEM) (Raju, Laffitte & Byrne, 2002; Stark, Chernyshenko & Drasgow, 2006). Of these approaches, while SEM can be used for examining measurement invariance at the test level, DIF can produce findings relating to item level invariance. IRT models are recommended for measurement invariance at the item level. However, the employment of more than one invariance determination technique together is recommended (Hambleton, 2006).

The different techniques used for DIF analyses may vary in terms of synchronization criteria, algorithms, and the cut-off point used to decide about the DIF status of a given item. Therefore, the results obtained from different DIF detection techniques cannot be interpreted to be in full agreement (Atalay, Gök, Kelecioğlu, & Arslan, 2012; Çepni, 2011; Gök, Kelecioğlu & Doğan, 2010). Accordingly, this study used IRTLR and SIBTEST techniques, which are among techniques for detecting DIF (Camilli & Shephard, 1994; Gierl, Khaliq & Boughton, 1999; Shealy & Stout, 1993). Also, HGLM technique, which was proposed by Kamata (2001) to examine measurement invariance for hierarchical and nested data and which was accepted to be advantageous, was also preferred (Pan, 2008; Rawls, 2009).

As a result of the measurement invariance analyses based on statistical techniques, it is possible to obtain results regarding the significance of a systematic difference between subgroups in which invariance is examined; yet, the cause of the difference cannot be interpreted (Osterlind & Everson, 2009). Accordingly, the difference between groups may arise from a real difference or item bias (Zumbo, 1999). The difference determined based on statistical techniques cannot be interpreted as an advantage or bias to a group. In cases where metric invariance cannot be achieved using SEM and items showing DIF are identified, the existence of item bias is suspected. Therefore, in cases where measurement invariance cannot be achieved with statistical methods, item bias studies (such as content analysis, expert opinion, empirical evaluations) are absolutely necessary to reveal the reasons for this situation (Zumbo, 2007). This study is evaluated to be highly significant in terms of carrying out item bias studies and determining the causes of DIF for items in the PISA 2015 science literacy test, which, if found any, are determined to show DIF as a result of DIF analyses done using culture and language variables.

There are several studies examining the measurement invariance related to PISA. For example, Kankaras and Moors (2014) examined the measurement invariance of PISA 2009 science literacy items with IRT based DIF detection techniques according to different countries. The presence of DIF in PISA 2006 science literacy items over the samples of Australia, Britain, and Turkey was discussed in another study (Başusta, 2013). Le (2006) examined the presence of DIF in science literacy items by country, language and gender groups according to the initial results for PISA 2006. The examination of the studies indicated that they addressed an item-level measurement invariance for PISA science literacy. On the other hand, in a study examining the test- level measurement invariance, the results were obtained over the data of two countries. In the present study, analyses were conducted and the results were obtained over the data of four countries to address the language and culture variables together. Besides, this study addressed test and item level measurement invariance together, and findings obtained based on expert opinions in addition to statistical analyses were supported with a bias study. With this respect, the study is different from other studies conducted on the PISA science literacy test. Also, no measurement invariance study had been conducted on PISA 2015 science literacy test before.

International education studies have important impacts not only on countries' education systems, but also concordantly on development levels. PISA is the largest of the international educational surveys that offer important outputs in terms of accountability in education. However, it is necessary that the comments and comparisons should be appropriate and the measurement instruments should be tested in terms of cross-cultural measurement invariance. Analyses regarding the measurement invariance of the measurement instruments are carried out in test and item levels. SEM is the most common and recommended method for examining test-level measurement invariance. With this method, it is possible to examine whether the test has the same factor construct across different groups. Also, the item-level examinations must be carried out together with test-level analyses, and two aspects must be evaluated together. The use of IRT-based methods is widely recommended for the item-level

measurement invariance analyses, namely DIF analyses. Moreover, hierarchical linear modelling is especially recommended for carrying out DIF analyses for hierarchical and nested data. Because methods employed for DIF analysis use different cut-off points and algorithms, use of more than one method and comparison of the results is definitely recommended. Accordingly, in this study, in addition to IRTLR and HGLM techniques, non-parametric SIBTEST technique was also preferred for DIF analysis. Findings in which measurement invariance across groups could not be statistically provided with DIF or SEM analyses can be obtained; yet, no causality has been mentioned for this condition. Therefore, findings obtained in cases where measurement invariance cannot be achieved should be supported with item bias studies. Bias studies can help evaluate whether the difference found through statistical analyses is a real difference or bias, and if it is a bias, they can help do evaluations about where it comes from. As the focus of the PISA 2015 application was the field of science, the problem of the study consisted of examining the cross-cultural measurement invariance based on different and same language forms of the science literacy test by using SEM, IRTLR, SIBTEST, and HGLM and if invariance was not achieved, determining the cause of the failure with the help of bias studies.

METHOD

Research Model

This study used a descriptive design since it aimed to examine test and item-level cross cultural measurement invariance and item bias of PISA-2015 science literacy test. Descriptive studies aim to reveal a case as it exists (Fraenkel & Wallen, 2006).

Research Group

The sample of the PISA 2015 application involved approximately 540 thousand 15-year-old student group selected by the stratified sampling method (OECD, 2016). The research group of this study was established by considering the cultural and language elements of the countries. Accordingly, Australia (English) and France (French) taking the tests in their native languages and Singapore (English) and Turkey (Turkish) taking the test in adapted languages were involved in the sample of the study. Besides, the inclusion of Singapore and Australia in the sample made it possible to make a similar language and different culture comparisons. Thus, the elements of culture and language were examined together.

Students from the four countries included in the study group taking the S12 booklet involving mainly science items were included in the study. However, the number of students taking the related item set varied by country. In the measurement invariance analyses, in which comparisons between models are based on examining the model fit indexes, the presence of different numbers of students in different groups may affect the results. Many studies have shown that sample size affects model fit indexes (Fan & Sivo, 2007; Mahler, 2011). For this reason, an equal number of students taking the S12 booklet, which is mainly made up of science items set, from each country sample were included in the study group. The distribution of students according to each country sample is given in Table 1.

As is seen in Table 1, the lowest number of students was in Singapore sampling with 556 students, followed by Turkey with 612 students, France with 624 students, and Australia with 1352. For the analysis, students as much as the number of students in Singapore sampling (556), which had the fewest number of students, were selected from other countries using the random sampling method.

Data Collection

The data of the study were obtained from the official website of the OECD at <www.pisa.oecd.org>. In the PISA 2015 application, which used computer-based assessments for the first time, in addition to computer-based assessment, paper-pen assessments were also employed for countries that preferred this method. In the PISA application, items found in the previous applications and the newly-developed items in each cycle are located in the test together. The items developed for the PISA 2015 were developed in accordance with computer-based assessments. Accordingly, in the PISA 2015, a total of 18 science-based item groups (12 computer-based, 6 paper-pencil-based) were evaluated. Also, released items are needed to conduct item bias studies. Given the released items and their item groups, the highest number of items released was from the S12 booklet. Accordingly, the responses of students who took the related item set from the four countries were included in the analysis. When the student responses were examined, items coded as multiple responses, unreachable, and not responded were coded as inaccurate and replaced by "0". This change was made to meet the dual scoring assumption for measurement invariance analyses. Again, for the same purpose, the partial scores of item S637Q02S in the item set S12, which is partially scored, were re-coded as '1', as a correct response. In the related item set, 17 of the 18 items were double-scored (OECD, 2017).

The factor structure of the test should be determined so that the measurement invariance can be examined in terms of the culture variable. Accordingly, when determining the factor structure of the test, sub-dimensions of cognitive levels

Table 1. Student distribution regarding countries in the sample

Country	Student size (included in the study)		Total student size	
	f	%	f	%
Australia	556	25	1352	43.0
France	556	25	624	19.8
Singapore	556	25	556	17.7
Turkey	556	25	612	19.5
Overall Total	2224	100	3144	100

related to science literacy were taken into consideration (OECD, 2017). The items in the related item set were distributed to sub-dimensions such as "Distinguishing scientific situations" (4 items), "Using scientific evidence" (6 items), and "explaining the facts scientifically" (7 items).

In this study, the item bias detection process was carried out over items that were released and accepted as showing DIF according to Turkey-Australia and Turkey- Singapore bilateral comparisons in which Turkish and English forms of the test were compared. In this sense, an expert opinion form was developed by the researcher to collect expert opinions. In the form, the forms of the items belonging to different cultures were included together. In the expert opinion form, experts were asked to report whether any items advantaged a culture group, and if it did, they were asked to indicate and explain the direction of the bias. The possible reasons for bias were presented to the experts' opinion with five items. In addition to the possible reasons listed, a blank was allocated on the form for other reasons that the expert wanted to specify. The experts separately expressed their opinions for five items showing DIF. Within the scope of the study, a total of 16 experts were consulted, including eight measurement and evaluation specialists who completed their doctorate education, two foreign languages instructors (one with Ph.D. and the other with master's degree), and six science instructors (four with Ph.D., and two with master's degree).

Data Analysis

Before analysis, the data were screened for extreme values and missing data. No individuals with extreme values were found, and individuals with missing values were excluded. As a result of necessary corrections, MG-CFA, GADM, MTK-OO and SIBTEST analyses were performed after assumption checks were made for each analysis.Before MG-CFA, the data were checked for normality and multicollinearity assumptions (Tabachnick & Fidell, 2007). Accordingly, the data set was found to meet the related assumptions. In the measurement instruments in PISA application, which employs measurement tools that are used in many countries, are adapted to different cultures. Before examining the measurement invariance or DIF existence in the items in such measuring tools according to groups (cultures), the invariance of the factor structure of the desired feature to be measured should be examined for each group separately (Sireci & Swaminathan, 1996). Accordingly, the factor structure for the science literacy test was formed according to cognitive levels, and the model fit for the related measurement model for each country was tested separately first. Model fit was evaluated according to the goodness of fit indices. In this evaluation, $X^2/sd \leq 2$, RMSEA ≤ 0.05, CFI ≥ 0.95, GFI ≥ 0.90, NNFI ≥ 0.95 and SRMR ≤ 0.05 criteria were considered. The first step in measurement invariance analysis with MG-CFA, which is based on testing four nested hierarchical models, is called structural invariance. In this step, the free inter-group estimation of error variances, regression constants, and factor loads are achieved, while the load pattern and number of factors are limited. By examining the fit of this model called "Model A", evaluations about structural invariance can be made. The second step that allows the limitation of inter-group factor loads and free estimation of factor loads, regression constants, and error variances is called metric invariance. The model established for this step is called "Model B". Metric invariance is based on the examination of model fit changes between Model A and Model B and the change in Chi-square (X^2) and CFI values. When examining the change in chi-square value, the critical X^2 value is compared with the difference between X^2 values of Model A and Model B according to the degree of freedom determined according to the difference between the degrees of freedom of the two models. Accordingly, if the X^2 value based on the difference is determined significant, metric invariance is considered to be provided. Since the X^2 value is sensitive to the sample size, the examination of the change in CFI value of the other goodness of fit indices is another recommended method. Accordingly, it is considered that metric invariance is provided if the difference between CFI values of Models A and B is in the range of $-0.01 \leq \Delta CFI \leq 0.01$. Thus, the inter-group invariance of factor loads can be mentioned. If metric invariance cannot be achieved, this is called weak invariance. If weak invariance only can be achieved, then bias is suspected. The next step after metric invariance is called strong invariance. In addition to other constraints in strong invariance investigations (factor pattern and factor load), regression constants are also limited. In the solid invariance, which is the last step, in addition to the previous step, error variances are also limited between the groups, and the significance of the change in the model fit is examined. To accept the inter-group comparisons as valid, the measurement tool should provide the measurement invariance at the least strong invariance level (Tabachnick & Fidell, 2007; Vandenberg & Lance, 2000). LISREL 8.8 program was used to conduct MG-CFA.

After the test-level measurement invariance analyses were carried out, item-level measurement invariance analyses were initiated. For the HGLM analysis performed for this purpose, homogeneity of variance, multicollinearity, normality of level-1 and level-2 errors, the analysis of the independence of errors, which are the assumptions of the analysis, were performed. Accordingly, the data set was found to meet HGLM assumptions (Raudenbush & Bryk, 2002). HLM 7 software was used for HGLM analyses. The IRTLR test technique was used for DIF detection analyses. For this purpose, first of all, the data set was examined according to the assumptions of local independence and unidimensionality, and both assumptions were found to be met (Embretson & Reise, 2000). Also, when the model was examined in terms of data compatibility, the data set was found to fit the 3-parameter model. IRTLRDIF software was used for IRTLR analyses. The SIBTEST technique, which is a non-parametric technique based on IRT, does not have any assumptions in addition to the other techniques, so the analysis was initiated directly (Shealy & Stout, 1993). SIBTEST 1.7 was used for the analysis. When determining focus and reference groups for DIF analyses, the country which was more successful in terms of PISA 2015 results was identified as the focus group in the related comparison. For example, in Singapore-Turkey comparison, Singapore was determined

as the focal group. The significance level was accepted as 0.05. Content analysis was used for expert opinions obtained within the context of item bias determination studies.

RESULTS

According to the results of the analysis, the findings and interpretations were discussed as follows: firstly, the findings and interpretations of the MG-CFA which was performed to examine the measurement invariance of the factor structure of the test in terms of culture variable were addressed. Next, findings and comments on IRT-LR, SIBTEST, and HGLM analyses performed to examine the measurement invariance of the test items in different cultures were included. Finally, the results of the bias study of the items that were released and determined to not show invariance as a result of the measurement invariance analyses at the item level were discussed in this section.

PISA 2015 Science Literacy Test Findings Related to Measurement Invariance of Factor Structure in terms of Cultural Variable

Prior to the measurement invariance analysis, the three-factor model established for science literacy was tested for each country data set. Confirmatory Factor Analysis results for this purpose are given in Table 2.

Table 2 shows the goodness of fit indices of the measurement model established for science literacy for each country. When the relevant indices are examined, the indexes can be said to be at an acceptable level. Accordingly, the science literacy test shows a good fit with the model established according to cognitive levels for each country data. The results of MG-CFA performed to examine the test-level measurement invariance were included for structural invariance and metric invariance, respectively. To investigate the structural invariance, which is the first step of examining the measurement invariance, the limitation of the factor number and factor pattern for each group, factor loads, regression constants, and goodness of fit values for Model A based on the free estimation of error variances are shown in Table 3 (Vandenberg & Lance, 2000).

The goodness of fit values of Model A established for the test of structural invariance in Table 3 were within acceptable ranges. In other words, the science literacy test could be interpreted to show structural invariance across different cultural groups. To make a decision about metric invariance, the change in model fit between "Model B" established by limiting factor loads and "Model A" with no parameter limitation was examined. Accordingly, the following calculations were made according to the indices in Table 3: $\Delta X^2 = 2085.19$ and $\Delta sd = 150$. Thus, the critical X^2value was determined as $X^2_{(150,0.05)} = 179.58$, where the degree of freedom was 150. The ΔX^2 value was found to be 2085.19> 179.58 when compared to the critical X^2 value. In this case, with inter-group factor loads limitation, the model fit was found to differ significantly. As is seen in Table 3, the change between Model A and B was 0.19. That the change in CFI value was not within the range of $-0.01 \leq \Delta CFI \leq 0.01$ showed a significant difference with the limitation of the model fit factor loads, similar to the change in X^2 value. This meant that the PISA 2015 science literacy test did not show metric invariance across different cultures. Since the invariance analyses consist of hierarchical steps, strong and strict invariance analyses, which are the next steps, could not be initiated. If there is evidence that metric invariance is not achieved in measurement invariance studies, this is accepted as an indicator of bias in items (Johnson, 1998). Therefore, it turned out that there was an item bias suspicion in the science literacy test examined in the scope of the study.

Metric invariance analyses were repeated with dual and triple combinations to determine which two or three of the Australia, France, Singapore, and Turkey samples included in the study caused the failure to achieve measurement invariance. First, factor loads were freed for each country separately and the metric invariance analyses were repeated for the remaining three countries. The results of the metric invariance analyses for the triple combinations of the countries are given in Table 4.

When the $\Delta\chi 2$ values in Table 4 were compared with critical chi-square value $X^2_{(33,0.05)} = 47.40$, which was determined according to the degree of freedom $\Delta sd=500-467=33$, $\Delta\chi 2$ was found to be $\Delta\chi 2 > 47.40$ for each triple group. Accordingly, it can be concluded that the change in model fit was significant when factor loads were limited. It can be said that the science literacy test did not show metric invariance

Table 2. Goodness of fit values relating to the model established for the science literacy test

Country	Statistics								
	X^2	Sd	X^2/sd	RMSEA	CFI	GFI	SRMR	AGFI	NNFI
Australia	100.74	101	1.00	0.0	1.00	0.98	0.029	0.97	1.00
France	136.85	101	1.36	0.025	0.99	0.97	0.035	0.96	0.99
Singapore	156.00	101	1.54	0.031	0.98	0.97	0.037	0.95	0.98
Turkey	129.68	101	1.28	0.023	0.96	0.97	0.037	0.96	0.96

Table 3. MGCFA results for measurement invariance of science literacy test according to culture variable

Model	X^2	sd	X^2/sd	GFI	RMSEA	CFI	NNFI	SRMR	ΔX^2 (Δsd)	ΔCFI
A	852.26	467	1.82	0.95	0.039	0.96	0.96	0.076	2085.19 (150)	0.19
B	2937.45	617	4.76	0.82	0.082	0.77	0.80	0.17		

Table 4. The results of the metric invariance analyses for the triple combinations of the countries

	χ^2	Sd	$\Delta\chi^2$	Δsd	CFI	ΔCFI
AUS-FRA-SGP	926.488	500	74.23	33	0.96	0.00
AUS-FRA-TUR	1138.648	500	286.39	33	0.94	0.02
AUS-SGP-TUR	1208.541	500	356.28	33	0.94	0.02
FRA-SGP-TUR	1176.210	500	323.95	33	0.94	0.02

Table 5. Metric invariance analyses for the dual combinations of the countries

	χ^2	Sd	$\Delta\chi^2$	Δsd	CFI	ΔCFI
AUS-FRA	887.252	484	35	17	0.96	0.00
AUS-SGP	890.429	484	38	17	0.96	0.00
AUS-TUR	1108.668	484	256	17	0.94	0.02
FRA-SGP	901.763	484	50	17	0.96	0.00
FRA-TUR	1017.117	484	165	17	0.95	0.01
SGP-TUR	1155.984	484	304	17	0.94	0.02

for four countries and that this did not stem from a single country. When $\Delta\chi2$ values were examined, the values for combinations involving Turkey sample, which was the only country taking the translated test form, were determined to be quite high compared to other triple combinations not involving Turkey. Another criterion is the change in CFI values in Table 4. Accordingly, ΔCFI relating to the model established by freeing the factor loads of Turkey sample and limiting the factor loads for the Australia-France-Singapore samples was determined to be 0.00. This value was found to be in the critical range. Unlike the change in Chi-square value, this can let us conclude that the model fit did not show a significant difference, that is, metric invariance was achieved for the other three countries through freeing the factor loads of Turkey sample. In triple combinations not involving Australia, France, or Singapore, ΔCFI was determined to be 0.02. Accordingly, this finding can lead to the interpretation that even if the factor loadings of these countries were freed, metric invariance was not achieved, that is, these countries were not the cause of failure to achieve metric invariance. If only the change in CFI value was considered, then the hindrance created by the Turkey sample, which took the translated version of the test, against achieving the measurement invariance may lead to the interpretation that there was an important distinction between the translated and original form of the test. Metric invariance analyses were repeated on dual groups among the four countries. Thus, to determine the countries that caused the failure to achieve metric invariance, the factor loads of two countries were freed and analyses were conducted on the other two countries. The results of the analysis are given in Table 5.

The examination of $\Delta\chi2$ values in Table 5 indicated that $\Delta\chi2 > 27.59$ when compared to $X^2_{(17,0.05)}$ critical Chi-square value according to Δsd = 484-467 = 17 degree of freedom. Accordingly, freeing the factor loads for dual groups did not change the failure to achieve metric invariance. However, the difference between $\Delta\chi2$ values belonging to groups not involving Turkey samples ($\Delta\chi2$ = 35, 38, 50) and $\Delta\chi2$ values of the groups involving Turkey ($\Delta\chi2$ = 165, 256, 304) and the fact that values were close to each other may give rise to the interpretation that the Turkey sample affected the model fit much more than other countries. The examination of the change in CFI values, one of the model fit indexes, indicated that ΔCFI value calculated to be 0.02 for dual groups involving Turkey (AUS-TR and SGP-TR) was not in the critical range, which meant that the difference in the model fit was significant, and the metric invariance was not achieved. The determination of ΔCF value calculated for the dual group involving France and Turkey samples as 0.01 may indicate that, although this value was at the upper limit of the critical range, the model fit did not change significantly, in other words, metric invariance was achieved. On the other hand, ΔCF values calculated for dual groups not involving Turkey (AUS-FRA, AUS-SPG, FRA-SPG) were determined as 0.00. This value was in the critical range, and this indicated that the model fit did not differ significantly and that metric invariance was achieved. Accordingly, similar to the change in the Chi-square value, changes regarding the CFI value in the groups involving the Turkey sample were higher.

There were no similarities between the countries in terms of the relationships between the responses to the items and the related factors. This may be interpreted that making a comparison between countries according to the scores obtained from the test was not meaningful. The model established according to the cognitive levels of science literacy was found to show structural invariance across different cultures. Accordingly, the source of the differences observed between the groups by scores obtained from the test may be considered as the measurement tool. Therefore, making a comparison between groups may not be correct. As a result of the results obtained, it is possible to say that the invariance between countries was a weak invariance. This was thought to have stemmed from various translation problems and cultural differences. It may also be indicative of a possible source of Differential Item Functioning (DIF) in the items.

PISA 2015 Science Literacy Test Findings Related to Measurement Invariance of Test Items in terms of Cultural Variable

The presence of DIF in the items of the science literacy test across cultures was investigated by HGLM, SIBTEST, and IRT-LR techniques. Dual combinations of cultures were established for the analyses (Australia-Singapore, Australia-France, Australia-Turkey, Turkey-Singapore, France-Singapore, and France-Turkey). Items showing DIF according to all three techniques and at least at the B-level were accepted to have DIF. The results regarding DIF analyses are given in Table 6.

According to the results of the DIF analysis in Table 6, 6 (35%) of the total 17 items were considered to show DIF as a result of the comparisons between Australia-Singapore and Australia-France. It is noteworthy that the items showing

Table 6. DIF analyses regarding the science literacy test items

Item no	Items showing DIF and the advantaged country					
	Singapore-Australia	**Australia-France**	**Australia-Turkey**	**Singapore-Turkey**	**France-Turkey**	**Singapore-France**
1	Singapore	Australia	Turkey	Turkey	France	-
2	Australia	-	-	Turkey	-	France
3*	-	-	-	-	-	-
4	Australia	Australia	Australia	-	-	-
5	Australia	-	Turkey	-	Turkey	-
6	-	-	-	Singapore	-	-
7*	-	-	-	-	-	-
8*	-	-	-	-	-	-
9	Singapore	France		-	-	-
10	-	-	-	-	-	Singapore
11	-	Australia	-	Singapore	-	-
12*	-	-	-	-	-	-
13	-	France	-	-	France	-
14	Australia	Australia	-	-	-	-
15	-	-	Australia	-	-	-
16*	-	-	-	-	-	-
17*	-	-	-	-	-	-
Total	6(%35)	6(%35)	4(%24)	4(%24)	3(%18)	2(%12)

*shows measurement invariance

the most DIF were determined among the groups that take the tests in the source language such as Singapore, France, and Australia samples. In Singapore-Turkey and Australia-Turkey comparisons, 4 items (24%) were designated to show DIF. Also, 3 items (18%) in France-Turkey comparison and 2 items (12%) in France-Singapore comparison were found to show DIF. Besides, while more items were expected to show DIF in comparisons involving the Turkey sample, which took the translated form of the test, findings indicated that there were much fewer items showing DIF. The interpretation is that sources, except for the translation mistakes, were also influential in terms of sources of bias between cultures.

Another finding in Table 6 was the presence of items showing measurement invariance in terms of all comparisons. Accordingly, 6 (35%) of the 17 science items were not determined to show DIF for all paired comparisons, instead, they were found to show measurement invariance. Accordingly, items S601Q01, S610Q02, S626Q03, S626Q04, S637Q05, and S641Q03 were determined to show measurement invariance, and S641Q03, S601Q01, and S637Q05 were among the published items. Also, as is seen in Table 6, there was no item commonly showing DIF in all comparisons.

Bias Study of PISA-2015 Science Literacy Items which did not show Measurement Invariance According to Culture Variable

With DIF, it is possible to obtain findings regarding the statistical significance of the systematic difference between the groups of items. However, a bias study is needed to comment on whether this difference is due to item bias or a real difference. To this end, a bias study was carried out based on the expert opinions about items that were released and accepted to show DIF according to the Singapore-Turkey and Australia-Turkey comparisons. Thus, expert opinions were obtained to determine if an item provides advantages to a country and, if it does, the possible causes of these conditions. The opinions of a total of 16 experts including 10 measurement and evaluation experts (two experts in the field of science, two experts in foreign language education, two foreign language instructors, and four science educators) were consulted in two stages. The expert opinion form was developed by the researcher. In the expert opinion form, both country forms for items showing DIF were placed together. In the first stage, the experts were first asked whether an item provided an advantage to a culture group, and if so, what the possible reasons were. In the second stage, based on the findings obtained from the first stage, experts were asked to re-submit their opinions on the same form to reach a consensus on the views. Tables 7 and 8 show the findings obtained in accordance with expert opinions obtained from the second stage. First, the distribution of opinions regarding the provision of any advantage as reported by the experts was addressed in Table 7.

When the expert opinions about the items that were published and found to exhibit DIF were examined (Table 5), the number of experts stating that items S641Q01, S641Q02, S641Q04, S637Q01, and S601Q04 advantaged a culture

group were determined as 15, 16, 16, 15 and 16, respectively. Although the items were stated to generally advantage Australia and Singapore, the opinion that item S641Q02 advantaged Turkish students in some situations and Singaporean students in others was adopted by eight experts. On the other hand, Table 8 gives the distribution of the opinions of the experts who stated the items advantaged a culture group on the possible causes of bias for these items.

The experts expressed their views on possible sources of bias in the items together with their reasons. Accordingly, as shown in Table 8, nine experts found bias in item S641Q01, which is about meteorites and craters, due to the use of words or phrases with a different meaning in the item. In books and textbooks in Turkey, some of the meteorites entering the atmosphere are called "göktaşı" (aerolite), and some are called "meteorit" (meteorite). This kind of situation was stated to cause confusion in students. The meteor fall in Australia in the year when PISA 2015 application was administered and the presence of the world's largest meteor traces in Australia may have resulted in higher awareness and knowledge levels of students participating in the practice from Australia. This may be providing advantage on behalf of Australian students. According to an expert, weight and gravity topics are addressed together in the Turkish curriculum, but gravity is not handled as much as weight. This was thought to be a disadvantage for Turkish students. Also, 15 experts stated the translation-based differentiation of the items as the source of bias. Accordingly, although the English statement did not include the expression "self", the third option in the Turkish version was added the expression "self". Another issue relating to translation was that in one of the options "The meteoroid is attracted to the mass of Earth" was translated as "Göktaşı, Dünya'nın dönüşü tarafından çekilir" (The meteoroid is attracted by the mass of Earth), where 'tarafından' was translated as (by), which make understanding more difficult for students. Instead, the experts suggested that a translation such as "Dünya'nın dönüşü, göktaşına çekim kuvveti uygular" (Earth's rotation exerts a gravitational force to the meteorite) would be more favorable.

When the expert opinions in Table 8 were examined, another item found to show DIF was S641Q02. For this item, it was stated that the words or expressions used in the item were used in a different meaning. According to nine experts agreeing on this view, concepts such as "meteor, aerolite, meteorite' took place in the related sources at the same time in Turkey, and this could cause a conceptual confusion. The views relating to these causes were the same for question S641Q01, which used the same item stem. There were 14 experts who thought that the items differed based on the translation. According to 8 experts stating that the item provided an advantage to the Turkey sample, while "burn up" mean "burn", in the translated form it was stated as "destroyed by fire". This gives a considerable clue for the formation of fewer craters. It does not form a crater since it is destroyed. However, in the English form, the meaning "destroyed by fire" was not clear. Since the item S641Q04 had the same item stem as the two items mentioned above, there were similar views about it. The opinions differing from those of the other items were related to repetitive opinions for the same item stem.

When the expert opinions presented in Table 8 for the item S637Q01 were examined, the use of expressions or words in the item in different meaning was determined to be a source of bias by 12 experts. The familiarity of a cultural group with the item content was determined to be a source of bias by 4 experts. The provision of advantage to a culture

Table 7. Distribution of expert opinions on items found to show DIF advantaging any culture

	Items				
	S641Q01	**S641Q02**	**S641Q04**	**S637Q01**	**S601Q04**
Expert opinions	**f**	**f**	**f**	**f**	**f**
Does not provide advantage to a culture group	1	-	-	-	-
Provides advantage to a culture group	15	16	16	16	16
Australia	15	-	16	16	-
Singapore	15	16	-	-	16
Turkey	-	8	-	-	-

Table 8. Distribution of expert opinions on possible causes of bias for items found to show DIF

	Items					
Expert opinions	**S641Q01**	**S641Q02**	**S641Q04**	**S637Q01**	**S601Q04**	**Total**
	f	**f**	**f**	**f**	**f**	**f**
1. Use of phrases or words in the item in different meaning	9	9	9	12	10	49
2. Familiarity of a cultural group with the item content	15	13	13	4	13	58
3. Provision of advantage to a culture group due to the item format or formal characteristics	1	12	3	3	15	34
4. Existence of cultural differences in terms of the skills measured by the item	2	11	8	9	4	34
5. Translation-based differentiation of the items	15	14	6	-	12	47

group due to the item format or formal characteristics was found to be a source of bias by 3 experts. The existence of cultural differences in terms of the skills measured by the item was determined to be a source of bias by 9 experts. Accordingly, in terms of the measured skills, there was the opinion that although Turkish students learn to design experiments, they are not familiar with the content of items that require a critical approach to a fictionalized design, which measures high-level thinking skills. However, the radiation expression evokes the issue of radioactivity, which was considered as a disadvantage for Turkish students.

For item S601Q04, the view that the expressions or words in the item were used in a different meaning was found to be a source of bias by 10 experts. Accordingly, the opinion that the concept of sustainability was not a concept known by Turkish students was shared by another expert. In Singapore, an island country where the livelihood of most people is based on fish production, students are familiar with fish farming, and this was considered to be a cause of bias. According to nearly all of the experts, the item had long and complex expressions, it was not very suitable for the age and developmental characteristics of the students, and more than one skill was presented and measured in a single item in a complex way. On the other hand, the existence of cultural differences in terms of the skills measured by the item was accepted as a source of bias by four experts. However, according to the experts, there were also translation-bound problems, as well. The translated statements and phrases were observed to be complex. For example, it was necessary to read the statement beginning with "before it flows from pool to pool, filtered saline water..." more than once. The Turkish text was found to be very difficult to understand. The term "shellfish" was translated as a mollusk, but it was stated that the expression "shellfish" was a more accurate translation for this item. Also, the concept of "organisms" was translated as "living beings" in this item, fish are also living creatures. This causes ambiguity for Turkish students.

When the sources of bias indicated for all of the items handled in the bias study presented in Table 8 were examined, the most stated source of bias (58) was found to be the familiarity of a cultural group with the item content. However, other reasons of bias were respectively listed in descending order as follows: the differentiation of the words or expressions (39), the translation-bound differentiation (47), the skills measured by the item (34), and provision of advantage to a culture group due to the item format or formal characteristics (34).

DISCUSSION

MG-CFA was conducted to examine the measurement invariance of the factor structure of the PISA 2015 science literacy test in terms of culture variable. The study concluded that the invariance of the model established according to the cognitive levels of the science literacy items regarding the four countries including Turkey, Australia, Singapore, and France was achieved only in structural invariance. In the metric invariance stage, which is the next step, the invariance of the test can be determined as a result of evaluations done according to the changes emerging in the model fit resulting from the limitations of inter-group factor loads. According to the findings, the test was concluded to not show metric invariance across cultures, so metric invariance analyses were repeated with dual and triple combinations to determine in detail which country or countries caused the failure to achieve measurement invariance. When the significance of the change in the model fit was examined according to the chi-square goodness of fit index, it was concluded that metric invariance could not be established in dual or triple comparisons. However, the change in chi-square value in the comparison including Turkey sample was observed to be significantly greater than those of the other comparisons. Moreover, examining the change in CFI value to determine the significance of the change in model fit is another recommended criterion. Accordingly, as a result of the dual and triple comparisons, the metric invariance was established in dual or triple combinations that did not include the Turkey sample. This showed that Turkey sample taking translation form of the test posed an obstacle for the measurement invariance, so it was thought that there was an important distinction between the source language form and the translated form. Failure to achieve metric invariance as a result of MG-CFA analysis raised suspicion of bias for the items in the test. In this case, the measurement invariance of the items in the test was examined with DIF analyses, and bias studies were carried out on items with DIF. Ulutaş (2015) examined the measurement invariance of the PISA 2006 scientific literacy test in the samples of Turkey and the United States. In the study, analyses were carried out on booklets 1 and 5. Accordingly, while the booklet 1 did not show invariance in terms of factor loads and error variances, it exhibited invariance in terms of correlations between factors. The booklet 5 was determined to show invariance in terms of factor structure for two countries.

The study investigated the cross-cultural invariance of the PISA 2015 science literacy items by using three different DIF detection techniques including IRT-LR, HGLM and SIBTEST. According to the findings, the number of items found to show DIF was the highest (35%) according to the Australian-Singapore and Australia-France comparisons taking the tests in the source language, which is a noteworthy outcome. Besides, compared to other countries, no significant difference was found in comparisons which included the Turkey sample, which was the only country that took the translated form of the test. As a result of the dual comparisons made for different country groups, an average of 25% of the items were found to show DIF. The items which were considered to show DIF were observed to generally show similarities between comparisons. Although the items identified as having DIF were found to indicate similarity between groups, there was no item found to commonly show DIF for all comparisons. While the items with DIF showed an advantage on behalf of Australia (10), there were an equal number of items on behalf of Turkey (5), France (5), and Singapore (5). Moreover, 35% (6 items) of the science literacy test items showed cross-cultural measurement invariance. Three of these items were among the published items.

Ulutaş (2015), who examined the PISA 2006 science literacy booklet 1 and 5 for the Turkish and US students in terms of showing DIF, determined that 16 (28%) of the items in booklet 1, and 24 (40%) items in booklet 5 showed DIF. Of these items, 25 were found to provide an advantage for US students and 15 for Turkish students. Kankaras and Moors (2014) examined the measurement invariance of PISA 2009 science literacy items with IRT-based DIF detection techniques according to different countries. Accordingly, 15 of the 17 items were determined to show DIF. When these items were examined, they were found to be mostly in favor of Southeast Asian countries. In another study investigating the PISA 2006 science literacy items in terms of bias by culture variable, the analyses were carried out on Canada, Australia, Britain, and Turkey samples (Başusta, 2013). Le (2006) examined the status of science literacy items in terms of DIF, by country, language, and gender groups according to the preliminary data of PISA 2006. According to the results obtained, the rate of items showing DIF was 10% for the gender variable, 25% for the country variable, and 39-59% for the language variable. It is noteworthy that the largest ratio was found to be related to the language variable. Uzun and Gelbal (2017) carried out a DIF and bias study on PISA 2006 scientific literacy test items including Turkey, Australia, England, and Canada subgroups. According to the study, as the linguistic and cultural differences increased, the number of items showing DMF increased, as well.

According to the bias study conducted on the items which were published and determined to show DIF as a result of the comparisons between Singapore-Turkey and Australia-Turkey, the experts (16 experts) stated that the items provided an advantage to a certain country group and that the cause of this bias was listed from the most prevalent to the least as follows: the familiarity of a cultural group with the item content, use of phrases or words in the item in different meaning, and translation-based differentiation of the items. Besides, the existence of cultural differences in terms of the skills measured by the item and provision of advantage to a culture group due to the item format or formal characteristics were stated to be an equal level of bias sources. In a study investigating the bias sources of PISA 2006 science literacy items according to Turkish and US students, Ulutaş (2015) determined the bias source of 9 items identified to show DMF based on expert opinions as "the familiarity of a cultural group with the item content". In a study investigating the bias in the PISA 2006 science items for different cultures and languages, Başusta (2013) determined that according to expert opinions, the source of bias came from differences in program, culture, and language. The bias sources of PISA 2006 science literacy items according to different country and language groups were found to come from differences in test translation, and cultural and educational programs (Le, 2006). According to another study examining the possible causes of this situation for the items showing DIF among the published items for the English and Turkish forms of the PISA 2006 science literacy test, differences in culture, language, translation, and program were among the possible causes of bias (Uzun & Gelbal, 2017). In many studies examining cultural and linguistic bias reasons of tests, bias reasons similar to those of this study were determined (Gierl & Khaliq, 2001; Yıldırım & Berberoğlu, 2009). In a study on the translation process of the PISA 2000 application, the average sentence and word lengths were found to vary for different languages during the adaptation process, and this situation was determined to be not controlled by individuals doing the translation (Grisay, 2003).

CONCLUSION

According to the findings obtained from the study, the suggestions for the practitioners and researchers are discussed under this title. In international studies using tests adapted to many cultures such as PISA, the practitioners are recommended to examine the measurement invariance of items and tests in terms of many variables to make fair and in-place comparisons according to the scores obtained from the tests. Therefore, in such applications, DIF detection analyses should be carried out meticulously on the scores obtained from the pilot application. Bias sources should be determined and necessary arrangements and corrections should be made according to the findings. One of the main reasons for bias determined in cross-cultural measurements is the fact that a culture group is familiar with the item content, therefore attention should be paid to this situation during the writing process of the items. Also, for PISA applications involving many culture-adapted tests, translation problems, another possible cause of bias, should be considered carefully. The participating countries should diligently conduct the selection of individuals preparing the national forms of the country involved and inform them of possible reasons for bias. If possible, training programs should be organized for this purpose.

According to the findings of the study, the recommendations for the researchers can be listed as follows: The rate of items showing DIF was found to be higher especially in comparisons made according to countries taking the original language form of the test (Singapore-Australia and Australia-France) than others. Studies can be carried out to determine the possible causes of this case, in other words, the causes of item bias in these groups and possible causes of bias. In this study, analyses were carried out based on the data obtained from the item set no S12. Similar studies may be conducted for other sets of PISA applications. In this study, IRT-LR, HGLM, and SIBTEST statistical techniques were employed for DIF detection analyses, in other words, for investigating the measurement invariance at the item level. DIF analysis of PISA science literacy test items can be carried out with techniques different from those employed in this study. In this study, bias studies were conducted on English and Turkish forms. Similar studies can be conducted for other languages or cultures. While identifying experts for bias studies, it was not possible to find experts who were familiar with both cultures compared. Experts who are familiar with both cultures should be preferred when obtaining opinions about different forms of culture regarding bias studies.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington: American Psychological Association.

Atalay, K., Gök, B., Kelecioğlu, H. & Arsan, N. (2012). Değişen madde fonksiyonunun belirlemesinde kullanılan farklı yöntemlerin karşılaştırılması bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,43*, 270-281.Retrieved from https://dergipark.org.tr/tr/pub/hunefd/issue/7795/102030

Başusta, N. B. (2013). *Differential item functioning analysis of PISA 2006 science achievement test in terms of culture and language* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey

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

Çepni, Z. (2011). *Değişen madde fonksiyonlarının SIBTEST, Mantel Haenzsel, Lojistik Regresyon ve Madde Tepki Kuramı yöntemleriyle incelenmesi.* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Demirtaşlı, R. N. (2014). Öğrenme, öğretim ve değerlendirme arasındaki ilişkiler. In N. Demirtaşlı (Ed), *Eğitimde ölçme ve değerlendirme (3-29)*. Edge Akademi: Ankara

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Ercikan, K., Roth, W. M. & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teachers College Record, 117*, 1–28.

Fan, X. & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types, *Multivariate Behavioral Research, 42*(3), 509-529. doi:10.1080/00273170701382864.

Fraenkel, J. R. & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25*(4), 280-296. doi: 10.2307/1585851.

Gierl M. J. & Khaliq S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement; 38*(2), 164-187. doi: Gierl, M. H., Khaliq, S. N., & Boughton, K. (1999, June 7-11). *Gender differential item functioning in mathematics and science: Prevalence and policy implications.* In Annual Meeting of the Canadian Society for the Study of Education, Canada Retrieved from https://pdfs.semanticscholar.org/cca9/52d14ef16ce1dc1d07ce0caf11609680b85d.pdf

Gök, B., Kelecioğlu, H. & Doğan, N. (2010). Degişen madde fonksiyonunu belirlemede Mantel-Haenzsel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Egitim ve Bilim, 35*, 3-16. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/19/28

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225-240. doi:10.1191/0265532203lt254oa

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*, 182-188. doi:10.1097/01.mlr.0000245443.86671.c4

ITC (2005). *International test commission guidelines for test adaptation*. London: Author.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA-Nachrichten Spezial, 3,* 1-40. Retrieved from http://www.ssoar.info/ssoar/handle/document/49730

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93. Retrieved from www.jstor.org/stable/1435439

Kankaras, M. & Moors, G. B. D. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381-399. doi:10.1177/0022022113511297

Le, L. T. (2006, April 7-11). *Analysis of differential item functioning*. In Annual Meeting of American Educational Research Association in San Francisco. Retrieved from https://www.acer.org/files/analysis_of_dif.pdf

Mahler, C. (2011). *The effects of misspecifcation type and nuisance variables on the behaviors of population fit indices used in structural equation modeling.* B.A: The University of British Columbia.

Niemann, D., Martens, K. & Teltemann, J. (2017). PISA and its consequences: Shaping education policies through international comparisons. *European Journal of Education, 52*(2), 175-183. doi:10.1111/ejed.12220

OECD (2016). *PISA 2015 results (Volume I): Excellence and equity in education. PISA.*Paris: OECD Publications

OECD (2017). *PISA 2015 technical report*. Paris: OECD Publications. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks. CA: SAGE Publications.

Pan, T. (2008). *Using the multivariate multilevel logistic regression model todetectdif: a comparison with HGLM and Logistic Regression DIF detection methods* (Unpublished doctoral dissertation). Michigan State University, Michigan, USA.

Raju, N. S., Laffitte, L. J. & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 527-529. doi:10.1037//0021-9010.87.3.517

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models.* Newbury Park, CA: Sage.

Rawls, A. M. W. (2009). *The importance of test validity: An examinatıon of measurement ınvariance across subgroups on a reading test* (Unpublished doctoral dissertation). University of South Carolina, South Carolina, USA.

Shealy, R. T. & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P.

Holland & H. Wainer (Eds.), *Differential item functioning* (197–240). Hillsdale, NJ: Erlbaum.

Sireci, S. G. & Swaminathan, H. (1996, October). *Evaluating translation equivalence: So what's the big DIF?* In Annual Meeting of the NortheasternEducational Research Association, Ellenville, NY. Retrieved from https://files.eric.ed.gov/fulltext/ED428119.pdf

Sjøberg, S. (2015). PISA and global educational governance–A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(1), 111-127. doi:10.12973/eurasia.2015.1310a

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting differential ıtem functioning with comfirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292- 1306.doi:10.1037/0021-9010.91.6.1292.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. New York: Allyn and Bacon.

Ulutaş, 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. doi:10.14689/ejer.2015.58.3.

Uzun, N. B. & Gelbal, S. (2017). PISA fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi. *Kastamonu Eğitim Dergisi, 25*(6), 2427-2446. Retrieved from https://dergipark.org.tr/tr/download/article-file/363695.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69. doi:10.1177/109442810031002.

Yıldırım, H. H. & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108-121.doi:10.1080/15305050902880736.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics, Psychometrics, 26,* 45-79, The Netherlands: Elsevier Science B. V

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