
Alatlı, B. (2022). An investigation of cross-cultural measurement invariance and item bias of PISA 2018 reading skills items. *International Online Journal of Education and Teaching (IOJET)*, *9*(3). 1047-1073.

Received : 21.02.2022
Revised version received : 26.04.2022
Accepted : 29.04.2022

# AN INVESTIGATION OF CROSS-CULTURAL MEASUREMENT INVARIANCE AND ITEM BIAS OF PISA 2018 READING SKILLS ITEMS

*Research article*

Betül Alatlı https://orcid.org/ 0000-0003-2424-5937
Faculty of Education, Balıkesir University, 10010, Balıkesir, Turkey

betulkarakocalatli@gmail.com

Biodata:

Betül Alatlı has been working as an assistant professor for six years in the department of Educational Measurement and Evaluation. Her research interests are Test Development, Adaptive and Use, Measurement Invariance, Differential Item Functioning, Item Bias.

# AN INVESTIGATION OF CROSS-CULTURAL MEASUREMENT INVARIANCE AND ITEM BIAS OF PISA 2018 READING SKILLS ITEMS

Betül Alatlı

betulkarakocalatli@gmail.com

## Abstract

The aim of this study is to investigate the cross-cultural measurement invariance of the reading skills items of the PISA 2018 at test and item level. Another aim of the study is to determine the item bias for items that do not show cross-cultural measurement invariance in line with expert opinions. A survey model was used in the study. The study sample consisted of a total of 3192 students who answered the questions of the Rapa Nui unit of reading skills test from Australia, China, and Turkey samples. Multi-Group Confirmatory Factor Analysis was conducted to examine the cross-cultural measurement invariance of the one-dimensional factor structure of the Rapa Nui unit. Item Response Theory Likelihood Ratio Test, Mantel Haenszel, Simultaneous Item Bias Test and Logistic Regression were used to investigate test items for Differential Item Functioning (DIF). As a result, the Rapa Nui unit of PISA 2018 showed only structural invariance according to countries in this study. The majority of the items showing differential item functioning were obtained from the comparison of Turkey and China. According to Australia-China and Australia-Turkey comparisons, five out of seven items were identified as showing DIF As a result of the bias study based on expert opinions, it was concluded that the items showed a bias in terms of the familiarity of a culture group with the content of the items, translation, the use of expressions in the item in different meanings, item format, and features measured by the item.

*Keywords:* PISA, reading skills, measurement invariance, differential item functioning, bias

## 1. Introduction

With the worldwide measurement and evaluation applications in education, countries obtain important outputs for their education systems and can guide the education reforms. In addition, the results obtained for many variables about education provide a picture of the education systems of countries, thereby giving an opportunity to examine the education system in a multi-faceted way. The main international educational assessments are Program for International Student Assessment- PISA, Trends in International Mathematics and Science Study –TIMSS, and Progress in International Reading Literacy Study - PIRLS. The PISA has been conducted every three years since 2000. PISA is applied to assess the mathematics and science literacy and reading skills of 15-year-old students who attend formal education. In addition, many other variables related to education are assessed through student, teacher, school, and parent questionnaires. PISA was last applied in 2021 and the major domain was determined as mathematical literacy. However, the data of the 2021 PISA application has not been shared yet. Therefore, the application whose latest data can be accessed is the PISA 2018, in which reading was the major domain. In the application, the

schools and 15-year-old students who attend formal education in the target country is selected by Organization for Economic Cooperation and Development (OECD) randomly. PISA, which was applied as a paper-and-pencil test until 2009, can also be applied online since 2012. More than 600,000 students from a total of 79 countries and economies, including 37 OECD countries, participated in the PISA 2018. It is remarkable that Turkey ranks second in terms of the highest increase in reading skills and the country with the highest increase in mathematics and science literacy performance. However, the fact that the difference in scores according to school type and regions is quite high in Turkey, in other words, it is in the 10th place among all countries in this sense is another remarkable result. In the PISA 2018, the top five countries in reading skills are B-S-J-Z China (Beijing, Shanghai, Jiangsu, and Zhejiang), Singapore, Macau China, Hong Kong China, and Estonia. Among these countries, only Estonia is a member of the OECD. Turkey, on the other hand, ranks 40th in all countries and 31st in OECD member countries. The average of all countries participating in PISA 2018 for reading skills performance is 453 and the average score of OECD countries is 487. In addition, the average score of B-S-J-Z China, which ranks first, for reading skills performance is 555, and the average score of Turkey is 466 (OECD, 2019). Thanks to international assessments such as PISA, the education systems of successful countries can be examined. Reforms to be implemented in education systems can also be guided by taking the education systems of successful countries as an example.

There are some centers carrying out translation, adaptation, implementation, analysis, and reporting procedures of PISA, which is conducted by the OECD, on a national scale. For example, the PISA is carried out by the Ministry of National Education in Turkey. PISA tests for each country are adapted to the culture of that country. In order for the results obtained from the tests used in international assessments and the inferences related to these results to be considered fair and appropriate, the assumption of measurement invariance must be met in terms of the related culture or language (Gierl, 2000; Reise, Widaman & Pugh, 1993; Vandenberg & Lance, 2000). The requirement of measurement invariance examinations for cross-cultural comparisons is highlighted by Test Adaptation Guidelines (International Test Commission-ITC, 2005) and Standards for Measurement in Education and Psychology (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), 1999). Accordingly, it is considered important to examine the measurement invariance of culture and language variables related to the PISA 2018. Since the major domain in the PISA 2018 is reading skills, a group of seven items in the "Rapa Nui" unit, which make up the reading skills items and also the published items, was examined within the scope of the research. Thus, it was aimed to carry out a bias study on the items that were published.

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 latent structure (AERA, APA & NCME,1999). Many statistical techniques have been developed to examine the measurement invariance of tests in terms of certain variables. The most widely used and recommended technique for test-level measurement invariance investigations is Multi-group Confirmatory Factor Analysis (MGCFA). Techniques for determining differential item functioning are frequently used to examine the item-level measurement invariance (Cheung & Rensvold, 2002; Little, 1997; Lord, 1980; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006; Vandenberg & Lance, 2000). The DIF determination techniques can be used to examine the probability of answering a given item correctly or whether the success rate differs among individuals who are in the subgroups based on the same skill level but have been assigned to

the group in terms of certain variables (Camilli & Shepard, 1994; Zumbo, 1999). When one group shows a better performance at all skill levels, then a uniform DIF occurs. However, if it gives an advantage to one group up to a certain skill level and in favor of the other group after a certain skill level, then a non-uniform DIF occurs (Swaminathan & Rogers, 1990). DIF determination techniques are based on two theories: Item Response Theory (IRT) and Classical Test Theory (CTT). DIF determination techniques differ in terms of algorithm, synchronization criterion, and cutoff point used to decide about DIF. However, it cannot be said that DIF determination techniques are completely compatible with each other. This situation is also demonstrated by research (Acar, 2008; Atalay, Gök, Kelecioğlu & Arslan, 2012; Bakan, Kalaycıoğlu, & Berberoğlu, 2010; Çepni, 2011; Doğan & Öğretmen, 2008; Gök, Kelecioğlu, & Doğan, 2010). For this reason, it is recommended that multiple DIF detection techniques should be used together (Hambleton, 2006). The likelihood ratio technique was chosen because it is based on the Item Response Theory (Camilli ve Shepard, 1994). Although SIBTEST is not a technique based on IRT, it estimates the true scores of individuals based on the answers given to items other than the item on which DIF analysis is made, and matches individuals according to these scores. However, the ease of application and interpretation, and the availability of the software required for the analysis were effective in choosing the SIBTEST technique as the DIF detection technique (Shealy ve Stout, 1993). The Mantel haenszel method is a practical and widely used DIF determination technique (Holland & Thayer, 1988; Millsap & Everson, 1993). Logistic regression can detect uniform and non-uniform DIF (Zumbo, 1999). Accordingly, this study was conducted to investigate whether the reading skills items of the PISA 2018 Rapa Nui unit showed differential item functioning according to the culture variable by using Likelihood Ratio, SIBTEST, Mantel Hanszel, and Logistic regression techniques. The likelihood ratio technique was used because it was based on the item-response theory.

Differential item functioning is used to examine the significance of a systematic difference in items in terms of subgroups. However, a causal explanation for the difference cannot be made (Osterlind & Everson, 2009). In this case, it is not known whether the difference between the groups is an actual difference or a difference due to an item bias (Zumbo, 1999). For this reason, it is recommended that item bias analyses should be carried out in cases where DIF occurs (Zumbo, 2007). Accordingly, in this study, it is considered important to examine the item bias regarding the PISA 2018 reading skills items, which were determined to show DIF.

There are several studies on measurement invariance of PISA reading skills. For example, Özmen (2014) conducted an item bias study to investigate whether the PISA 2009 reading skills items showed DIF in terms of Turkey-United Kingdom and Turkey-USA groups. In the study, a total of four techniques were used based on the comparison of SIBTEST, MH, IRT-LR, and b parameters. In the bias analysis conducted according to the items that were identified to show DIF, it was concluded that the items could show bias due to cultural differences, adaptation errors, item format, and difficulty in understanding words. The DIF status of the PISA 2000 and PISA 2001 reading skills items in terms of different language groups was examined, and it was determined that local dependence on the items belonging to the same reading text could cause DIF. At the same time, it was pointed out that even though translations were done appropriately, translation bias might have occurred (Grisay & Monseur, 2007). Asil and Brown (2016) examined the measurement invariance of the PISA 2009 reading skills test using the data of 55 countries, taking Australia as the reference group. Accordingly, it was determined that the socio-economic source of education played an important role in measurement invariance. On the other hand, language factors and

educational practice were found to play a smaller role in measurement invariance. Ceyhan (2019) compared the measurement invariance between different countries according to the measurement language of the PISA 2012 reading skills test based on the same language (English-English, French-French) and different countries. In addition, different language pairs, including Turkish-Mandarin, Turkish-Spanish, and Turkish-English, were compared. The same-languages comparison provided structural invariance, while weak invariance was obtained for different languages. Accordingly, it was concluded that comparisons between these groups would not be significant.

Söyler (2020) made comparisons in terms of English speaking countries and non-native English-speaking countries for the PISA 2015 reading skills subtest. Canada, the USA, and the UK were involved in the study as English-speaking countries and Japan, Thailand, and Turkey as non-native English-speaking countries. Accordingly, it was reported that the PISA 2015 reading skills test did not show measurement invariance in terms of the language variable and that it would not be correct to make comparisons between countries according to the results obtained from this test. The measurement invariance of the PISA 2009 reading skills test was examined according to Spanish, Basque, Galician, and Catalan languages. According to the comparisons made according to different languages for the application of the test in Spain, it was concluded that the reading skills test showed metric invariance (Oliden & Lizaso, 2013). There is a general acceptance that PISA application is based on equivalent criteria for all languages and countries and that student performance is reliable and valid for international comparisons. However, some studies indicate that many factors, such as translation, familiarity with item content and format, curriculum differences, conditions of application, and linguistic and cultural factors, can prevent the comparability of scores. This situation affects the validity of PISA tests and items (Elosua & López-Jaúregui, 2007; Grisay & Monseur, 2007; Hambleton, Merenda & Spielberger, 2005; He & van de Vijver, 2012; Kreiner & Christensen, 2014; Mazzeo & von Davier, 2008; Oliveri & von Davier, 2011; Walker, 2007; Wetzel & Carstensen, 2013). Grisay and Monseur (2007) claim that "when a test is translated, it will always lose some equivalence." It is emphasized that full equivalence can never be achieved between multilingual tests, that is, a high level of comparability cannot be achieved (Arffman, 2010). Elosua and Mujika (2013) examined the comparability of the PISA 2009 reading skills test in terms of different languages spoken in Spain (Spanish, Basque, Catalan, and Galician) and determined that it showed metric invariance. However, for PISA 2000 reading skills, there were fewer items showing DIF in different countries that speak the same language (e.g., New Zealand, Ireland, the USA) than in groups that speak different languages (e.g. Canadian-English/French, Swiss-German/French) in the same country (Grisay & Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009).

With the PISA application, many variables are assessed in addition to the evaluations of science and mathematics literacy and reading skills. In order for the interpretations made according to the results of this application, in which important outputs related to the education systems of the countries are obtained, to be comparable, the items or tests used should meet the measurement invariance assumption, especially in terms of culture or language variables. The MGCFA method, which is frequently used for test-level measurement invariance studies, is recommended. Thus, the invariance of factor structure between groups can be examined. In addition to test-level invariance, item-level invariance analyses are considered very important for intergroup comparisons. M-H, SIBTEST, IRT-LR, and LR techniques were used in this study, since more than one technique that is recommended for measuring item-level invariance, in other words, differential item functioning, is used together. However, the determination of DIF is not considered as stand-

alone proof that a given item shows a bias between groups. For this reason, it is recommended that the item bias study should be carried out on the items which show DIF. In this study, it is considered necessary to examine the measurement invariance in terms of culture variable at the test level for the reading skills items of the PISA 2018 “Rapa Nui” unit and at the item level for the seven items in this unit and to examine the bias in items that do not show measurement invariance.

## 2. **Method**

### 2.1. Research model

Since this research was conducted to examine the test-level and item-level measurement invariance and item bias of the reading skills items of PISA 2018 in terms of the culture variable, the survey model was employed. In survey model studies, it is aimed to reveal an existing situation as it exists (Fraenkel & Wallen, 2006; Karasar, 2011).

### 2.2. Study group

A total of 79 countries, including 37 OECD members, participated in the PISA 2018 application. More than 600,000 students from these countries, representing 32 million students in the 15-year-old group, participated in the test (MEB, 2019). The study sample consisted of Australia, where students took the test in the source language, which is English, and Turkey and China, where students took the adapted form of the tests. In the PISA 2018 application, unlike the previous applications, students were first given a small test with basic questions to determine their levels, and then they took the test suitable for their level, instead of taking a fixed test. Therefore, students took individualized tests. Since the purpose of this research is to examine the bias in the reading skills items of PISA 2018, the study should be carried over the published items. For this reason, the research was conducted on the answers given to the seven items in the unit coded "Rapa Nui". Table 1 shows the distribution of students who answered the questions of the "Rapa Nui" unit in the samples of Australia, Turkey, and China, and the students included in the study by country.

Table 1. Student distribution of the countries included in the sample

| Country | Number of Students (Included in the Study) | | Total Number of Students | |
|---|---|---|---|---|
| | f | % | f | % |
| Australia | 1064 | 33.33 | 2755 | 40.34 |
| Turkey | 1064 | 33.33 | 1064 | 15.58 |
| B-S-J-Z China | 1064 | 33.33 | 3010 | 44.08 |
| Total | 3192 | 100.00 | 6829 | 100.00 |

In the study, first of all, students who responded to the items of the Rapa Nui unit were determined. Then, since it is known that model fit indices are affected by sample size, it was aimed to include the same number of students from each country in the study (Hu & Bentler, 1998; Fan, Thompson & Wang, 1999; Lei & Lomax, 2005; Fan & Sivo, 2007; Mahler, 2011). For this reason, since the sample of Turkey included the least number of students (1064), the same number of students was chosen from other countries randomly. Accordingly, the data obtained from 3192 students were included in the study.

### 2.3. Data collection

There were two versions of the PISA 2018 application: online test and paper-pencil test. However, in 70 of the 79 countries participating in PISA 2018, the tests were computer-based. Turkey, B-S-J-Z China, and Australia were also among the countries taking the online version of the test. In the PISA 2018 application, individualized tests were used for the first time. Accordingly, students were first given some basic items which were prepared to determine their level, and then they were given a test according to their level. For this reason, there was no common booklet taken by students (OECD, 2019). This makes it possible to examine the invariance of the units in PISA for measurement invariance examinations at the test level. In addition, the data obtained from the published items were included in the research to examine the DIF effect and item bias in terms of culture and language variables. Accordingly, the data obtained from the seven items in the Rapa Nui unit were included in the scope of this research. Table 2 shows the distribution of the items by code, item type, and measured cognitive processes.

Table 2. Distribution of item codes, types, and cognitive processes

| Research code | Item Code | Item Type | Cognitive Process | Sub-Cognitive Process |
|---|---|---|---|---|
| M1 | CR551Q01 | Simple multiple choice | Access to Information | Scanning and finding information in the text |
| M5 | CR551Q05 | Open-ended | Understanding | Expressing the literal meaning |
| M6 | CR551Q06 | Complex multiple choice | Evaluation and reflection | Reflecting on the content and format of the text |
| M8 | CR551Q08 | Simple multiple choice | Access to Information | Scanning and finding information in the text |
| M9 | CR551Q09 | Simple multiple choice | Evaluation and reflection | Identifying and overcoming conflicts |
| M10 | CR551Q10 | Complex multiple choice | Understanding | Combining and making inferences |
| M11 | CR551Q11 | Open-ended | Evaluation and reflection | Identifying and overcoming conflicts |

(OECD, 2019)

As seen in Table 2, three of the items examined within the scope of the research are simple multiple-choice, two are complex multiple-choice, and two are open-ended. Full scores coded as “2” obtained from the item coded as "CR551Q06" were re-coded as “1” and partial scores coded as “1” were re-coded as “0”. All of the other items were coded as “1-0”, and no changes were made in these items. When the items were examined in terms of their cognitive processes, items M1 and M8 were specified as "scanning and finding information in the text", item M5 as "expressing the literal meaning", item M6 as "reflecting on the content and format of the text", items M9 and M11 as "identifying and overcoming conflicts”, and item M10 as “combining and making inferences” (OECD, 2019). Accordingly, it can be said that there is diversity in terms of cognitive characteristics measured.

Since this study was conducted to examine whether the reading skills items of the PISA 2018 showed item bias in terms of culture and language variables, experts were consulted about the items that were accepted to show DIF. For this purpose, an opinion form was developed by the researcher to obtain the views of the experts on item bias in terms of culture and language variables. Within the scope of the study, a total of seventeen experts, including nine measurement and evaluation specialists who have a PhD, one measurement and evaluation specialist who have a master's degree, two foreign language educators (who have a master's degree), one expert with a PhD from the field of education programs and teaching (who have an undergraduate degree in foreign languages), and three experts from the field of Turkish language teaching (two of whom have a PhD and one with a master's degree), were consulted.

### 2.4. Data analysis

Before the data analysis was initiated, the data set was examined in terms of missing data and extreme values. Individuals with missing data were excluded from the analysis, as the identified missing data were less than 5% of the total data (Tabachnick & Fidell, 2018). Accordingly, as a result of the examinations, the data were organized for analysis. The Mahalanobis distances that were examined to determine the multivariate extreme value ranged from 36.95 to 4.18. Accordingly, it was determined that there was no extreme value at 0.001 significance level according to Mahalanobis distances. The total unit score was obtained from the sum of the scores obtained from the items, and the assumption of normality was examined. As a result of the analyses, the skewness and kurtosis coefficients obtained over the total scores ranged between -1 and +1. Accordingly, it can be said that the assumption of normality was met (Cokluk, Şekercioğlu, & Büyüköztürk, 2018). Another assumption required for MGCFA is multicollinearity, which is shown as an indicator of the relationships between independent variables (Tabachnick & Fidell, 2018). The variance inflation factor (VIF) values obtained as a result of the analyses performed to test this assumption were examined, and they were found to range between "1.004" and "1.454". In cases where VIF values are less than 10, there is no multicollinearity. Tolerance values were between "0.688" and "0.996". In cases where tolerance values are greater than 0.01, there is no multicollinearity. Accordingly, it was observed that there was no multicollinearity between the variables according to both VIF and tolerance values. The assumptions for the MGCFA were examined, and it was found that the data set met the assumptions. It is recommended to examine test-level invariance first in the analysis of item-level invariance between groups (Sireci & Swaminathan, 1996). For this reason, to examine the cross-cultural invariance of the reading skills items of the PISA 2018, the model fit of the single-factor structure of the Rapa Nui unit for each culture group was also examined. The model fit was evaluated over the following criteria: "$X^2/sd \leq 2$, RMSEA $\leq 0.05$, CFI $\geq 0.95$, GFI $\geq 0.90$, NNFI $\geq 0.95$ and SRMR $\leq 0.05$". MGCFA is based on the examination of four nested hierarchical models. In the structural invariance analysis, which is the first step, while free estimation of regression constants, factor loadings, and error variances between groups is allowed, load pattern and the number of factors are limited. Model fit is interpreted by examining the model fit indices obtained as a result of structural invariance. In cases where structural invariance is provided, metric invariance analyses can be performed. In metric invariance analysis, which is the second step, factor loadings are limited between groups, while the free estimation of regression constants and error variances is allowed. The variance in model fit is evaluated by examining the difference between fit indices obtained in structural invariance and those obtained in metric invariance. For this purpose, the significance of $\Delta X^2$ value is examined in comparison to $\Delta sd$ value. If there is a significant

variance in the value of "$X^2$", it is interpreted that metric invariance is not achieved. Since the $X^2$ value is sensitive to sample size, it is recommended to examine the variance in CFI value, as well. Accordingly, if the difference (ΔCFI) between the CFI values obtained from the structural and metric invariance models is in the range of "$-0.01 \leq \Delta CFI \leq 0.01$", then metric invariance is achieved. In the third step, with the analysis of the variance in the model fit by limiting regression constants in strong invariance and error variances in strict invariance, interpretations about invariance can be made. It is recommended to provide strong invariance to make comparisons between groups. In cases where even metric invariance cannot be achieved, the suspicion of item bias arises (Brown, 2006; Cheung & Rensvold, 2000; Hu & Bentler, 1998; Şencan, 2005; Şimşek, 2007; Vandenberg & Lance, 2000; Tabacknick & Fidell, 2018; Wu, Li, & Zumbo, 2007). The LISREL 8.8 software package was used for MGCFA.

In the study, the DIF effects for culture variable in PISA 2018 reading skills items were examined by using MH, LR, IRT-LR, and SIBTEST techniques. Accordingly, exploratory factor analysis (EFA) was conducted for the data of each country to determine whether the data obtained from the responses to the items provided the unidimensionality assumption for IRT-LR, the DIF determination method that is based on IRT. The EFA results are given in Table 3 and the scree plot is presented in Figure 1.

Table 3. Exploratory Factor Analysis Results

| Country | Count of factors | Eigenvalues | Explained variance | Total variance explained |
|---|---|---|---|---|
| Turkey | 1 | 5.863 | 48.857 | 48.857 |
| | 2 | 1.310 | 10.917 | 59.774 |
| Australia | 1 | 2.251 | 32.151 | 32.151 |
| | 2 | 1.009 | 14.409 | 46.559 |
| China | 1 | 2.360 | 33.720 | 33.720 |
| | 2 | 0.948 | 13.545 | 47.266 |

As seen in Table 3, the first factor contributes much more to the variance than the second factor, as shown by the eigenvalue difference between the first component and the second component. In addition, scree plots were also examined. Accordingly, a sudden decrease in the curve after the first component and a plateau after the second component indicates that the data meet the unidimensionality assumption (Gierl, 2000; Hambleton & Swaminathan, 1989). Another assumption of the IRT is local independence. It is considered important to meet the unidimensionality assumption for local independence, which is defined as the independence of the responses given to each item, that is, the items of a test should not be associated with each other. It is suggested that when the unidimensionality assumption is met, the covariance between the responses of individuals with similar capacity to the items is zero, and therefore, the assumption of local independence is also met (Hambleton & Swaminathan, 1989; Hambleton, Swaminathan & Rogers, 1991). Another assumption regarding IRT is model-data fit. As a result of the model data fit analyses conducted for this assumption, the B-S-J-Z China and Turkey datasets conformed to the 2-parameter model, and the Australian dataset conformed to the 3-parameter model. The DIF level is determined by the "$G^2$" value, which is obtained by the IRT-LR determination technique. Accordingly, the following interpretations are made: $3.84 < G^2 < 9.4$, no DIF effect or negligible; $9.4 \leq G^2 < 41.9$, moderate; $G^2 \geq 41.9$, high level of DIF (Greer, 2004).

The Mantel-Haenszel technique is one of the DIF determination techniques for dichotomously scored items. The ΔMH statistics obtained with the M-H technique show the degree of DIF. Accordingly, the ΔMH is interpreted as follows: ΔMH| <1, no DIF effect or negligible (level A); 1< |ΔMH| <1.5, moderate (level B); |ΔMH| ≥ 1.5, a high level of DIF (level C) (Dorans & Holland, 1993). Due to the limitations of the MH method as it can determine only uniform DIF and due to the high probability of type 1 error, it is recommended to be used with different techniques. Another DIF technique used in this research is Logistic Regression (LR). The DIF level is decided according to the ΔR2 value obtained from the LR technique. Accordingly, the $\Delta R^2$ value is interpreted as follows: 0 < ΔR2< 0.035, no DIF or negligible; $0.0035 \leq \Delta R^2 < 0.07$, moderate level; $\Delta R^2 \geq 0.07$, a high level of DIF (Jodoin & Gierl, 2001). According to another source, there is no DIF or it is negligible if ΔR2< 0.13; it is moderate if $0.13 \leq \Delta R^2 < 0.26$, and there is a high-level of DIF if $\Delta R^2 \geq 0.26$ (Zumbo and Thomas, 1996). The SIBTEST technique, one of the DIF determination techniques used in the research, is non-parametric. The DIF level is decided according to the "β" value obtained as a result of SIBTEST. Accordingly, |β|<0.059 is interpreted as a negligible level of DIF; 0.059≤|β|<0.088, as moderate level; |β|≥0.088 as a high level of DIF (Gotzmann, Wright, & Rodden, 2006; Stout & Roussos, 1995). For DIF analysis, the "difR" package (Magis, Beland, & Raiche, 2016), the "mirt" package (Chalmers, 2018), and the "ltm" package (Rizopoulos, 2006) of the R software were used. According to the PISA 2018 results, the successful country group was determined as the focus group in the DIF analyses. Accordingly, China was determined as the focus group in the Australia-China comparison, China in the Turkey-China comparison, and Australia in the Australia-Turkey comparison. The frequency values of the expert opinions that were obtained to determine the item bias were analyzed and interpreted.

## 3. Findings

Findings and interpretations about whether the seven items in the Rapa Nui unit, which includes published items of the PISA 2018 reading skills items, show measurement invariance for culture variable, whether the items show a DIF effect and the bias analysis for items that are accepted to show a DIF effect are discussed under this heading. In line with the answers given to the items in the Rapa Nui unit that is examined within the scope of the research, a CFA was conducted to determine whether the one-dimensional structure of the unit was confirmed for the data of each country. The results of the analysis are given in Table 4.

Table 4. CFA Results for Rapa Nui Reading Skills Unit of PISA 2018

| | Statistics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | $X^2$ | Sd | $X^2$/sd | RMSEA | CFI | GFI | SRMR | AGFI | NNFI |
| Australia | 21.25 | 14 | 1.52 | 0.022 | 0.99 | 0.99 | 0.021 | 0.99 | 0.99 |
| China | 29.17 | 14 | 2.08 | 0.032 | 0.99 | 0.99 | 0.023 | 0.98 | 0.98 |
| Turkey | 31.53 | 14 | 2.25 | 0.034 | 0.99 | 0.99 | 0.024 | 0.98 | 0.99 |

As seen in Table 4, the goodness-of-fit indices are in good agreement or at an acceptable level according to the CFA results conducted to determine whether the one-dimensional structure was confirmed according to the items in the unit for each country (X2/sd ≤ 2, RMSEA ≤ 0.05, CFI ≥ 0.95, GFI ≥ 0.90, NNFI ≥ 0.95, and SRMR ≤ 0.05). The first step in the investigation of measurement invariance with MGCFA is to examine the structural invariance. To examine the structural invariance, the factor pattern and the number

of factors were limited by setting the error variances, regression constants, and factor loads free in each group (Vandenberg & Lance, 2000; Wu, Li & Zumbo, 2007). The results of the structural invariance analysis are given in Table 5.

Table 5. MGCFA results for measurement invariance of PISA 2018 Rapa Nui reading skills unit by culture variable

| Model | $X^2$ | sd | $X^2$/sd | GFI | RMSEA | CFI | NNFI | SRMR | $\Delta X^2$ (Δsd) | ΔCFI |
|---|---|---|---|---|---|---|---|---|---|---|
| Structural invariance | 81.95 | 42 | 1.95 | 0.99 | 0.030 | 0.99 | 0.98 | 0.024 | 325.51 (14) | 0.08 |
| Metric invariance | 407.46 | 56 | 7.28 | 0.96 | 0.077 | 0.91 | 0.90 | 0.092 | | |

As seen in Table 5, the goodness of fit indices related to structural invariance show a good fit. Accordingly, it can be interpreted that the Rapa Nui reading skills unit shows a cross-cultural structural invariance. Since invariance analyses with MGCFA are based on nested models, unlike structural invariance, factor loadings are limited between cultures for metric invariance analysis, which is the second step. With this limitation, it is possible to comment on metric invariance by examining the variance in model fit. Accordingly, when the indices in Table 5 were examined, the variance in the value of "χ2" was found as Δχ2=325.51, Δsd=14. The critical value of χ2 according to Δsd=14 (degree of freedom) was "$\chi^2_{(14,\ 0.05)} = 23.68$". Accordingly, since 325.51>23.68, it can be interpreted that a significant variance occurred in model fit when factor loads were limited. ΔCFI is another recommended value for examining the variance in model fit. If this value is in the range of "-0.01≤ΔCFI≤0.01", it can be interpreted that there is no significant variance in model fit. However, as seen in Table 5, ΔCFI was calculated as 0.08, and it was not in the related range. Accordingly, the Rapa Nui unit of the PISA 2018 did not show cross-cultural metric invariance. It is stated that in cases where metric invariance is not achieved, suspicion of item bias arises (Johnson, 1998; Prelow, Tien, Roosa, & Wood, 2000). Since the Rapa Nui unit did not yield metric invariance, metric invariance analyses were repeated by considering the samples of Australia, China, and Turkey in pairs. Thus, it was aimed to determine whether any of the countries affected the invariance. The results of the metric invariance analysis obtained accordingly are given in Table 6.

Table 6. The results of the metric invariance analysis for the Rapa Nui unit of the PISA 2018 in terms of two cultures

| | $\chi^2$ | Sd | Δχ2 | Δsd | CFI | ΔCFI |
|---|---|---|---|---|---|---|
| Australia-China | 219.45 | 49 | 188.01 | 7 | 0.96 | 0.03 |
| Turkey- China | 154.29 | 49 | 253.17 | 7 | 0.97 | 0.02 |
| Turkey- Australia | 320.90 | 49 | 86.56 | 7 | 0.93 | 0.06 |

Table 6 shows the findings regarding the variance in model fit for the other two cultures when the factor loadings for one of the cultures were set free. Accordingly, the examination of the variance in model fit indicated that for example, when factor loadings were set free for the Turkish sample and limited for Australia and China and when the variance in model fit was compared to the critical chi-square value $X^2_{(7,\ 0.05)} = 14.067$, metric invariance could not be achieved for Australia-China, either, since 188.01>14.067 and the ΔCFI=0.03 value was not in the "-0.01≤ΔCFI≤0.01" range. Similarly, since Δχ2=253.17>14.067 and ΔCFI=0.02 value was not in the range of "-0.01≤ΔCFI≤0.01" for the Turkey-China pair, it was determined that metric invariance was not provided for these two countries, either. Metric invariance could not be provided for the Turkey-Australia pair,

either, since $\Delta\chi2=86.56>14.067$ and $\Delta CFI=0.06$ value was not in the range of "$-0.01\leq\Delta CFI\leq0.01$". Accordingly, metric invariance could not be achieved for all cultures.

The status of the PISA 2018 Rapa Nui reading skills items for showing DIF effect by culture variable was examined by using the MH, SIBTEST, IRT-LR, and LR techniques and making comparisons between Australia-China, Australia-Turkey, and Turkey-China. The results of the DIF analysis for the Australia-China comparison are given in Table 7.

Table 7. The DIF Analysis Results of the Rapa Nui Unit Reading Skills Items of the PISA 2018 for Australia and B-S-J-Z China comparison

| | M-H Method | | SIBTEST | | IRT-LR | | Logistic | |
|---|---|---|---|---|---|---|---|---|
| | ΔMH | Level of effect | $X^2$ | Level of effect | $G^2$ | Level of effect | $X^2$ | Level of effect |
| M1 | 3.97 | A | 6.79 | B | 2.48 | - | 4.11 | - |
| M5 | 26.98 | C | 18.93 | C | 40.02 | B | 27.77 | A |
| M6 | 21.36 | B | 43.45 | C | 17.33 | B | 139.57 | B |
| M8 | 161.10 | C | 98.37 | C | 300.27 | C | 188.66 | B |
| M9 | 88.78 | C | 75.87 | C | 119.91 | C | 91.34 | A |
| M10 | 4.45 | A | 5.50 | A | 0.09 | - | 5.41 | - |
| M11 | 13.60 | A | 7.57 | B | 18.17 | B | 16.31 | A |

As seen in Table 7, according to the comparison of Australia and China, items M5, M6, M8, M9, and M11 of the PISA 2018 reading skills unit were determined to show a DIF effect in terms of M-H, SIBTEST, IRT-LR, and LR DIF determination techniques. The criterion for the items so that they could be accepted to show a DIF effect was considered as showing at least B level of DIF according to at least two DIF determination techniques. The item coded as M1 showed a negligible level of DIF according to M-H and SIBTEST techniques. The item coded as M10 showed only a moderate level of DIF according to the SIBTEST technique. Therefore, items coded as M1 and M10 were not accepted to show a DIF effect. When the items showing and not showing DIF were examined in terms of the culture variable, it was seen that they had different characteristics regarding cognitive processes and sub-cognitive processes (Table 7). However, five out of seven items showed DIF. In the comparison between Australia and B-S-J-Z China, all of the items considered to show DIF showed a DIF effect in favor of B-S-J-Z Chinese culture. The results of the Turkey-China comparison regarding whether the PISA 2018 Rapa Nui reading items showed a DIF effect by culture variable are given in Table 8.

Table 8. The DIF Analysis Results of the Rapa Nui Unit Reading Skills Items of the PISA 2018 for Turkey-China Comparison

| | M-H Method | | SIBTEST | | IRT-LR | | Logistic | |
|---|---|---|---|---|---|---|---|---|
| | ΔMH | Level of effect | $X^2$ | Level of effect | $G^2$ | Level of effect | $X^2$ | Level of effect |
| M1 | 31.49 | B | 41.89 | C | 26.83 | B | 36.02 | A |
| M5 | 20.81 | B | 7.72 | A | 34.19 | B | 20.84 | A |
| M6 | 207.83 | C | 230.33 | C | 326.63 | C | 244.76 | C |
| M8 | 19.48 | B | 4.28 | A | 54.15 | B | 26.42 | A |
| M9 | 66.10 | C | 48.35 | C | 91.28 | C | 69.35 | A |
| M10 | 64.30 | C | 53.36 | C | 104.24 | C | 67.49 | A |
| M11 | 15.20 | B | 21.72 | C | 10.93 | B | 25.85 | A |

As seen in Table 8, all seven items in the Rapa Nui unit showed DIF according to Turkey-China comparison in terms of M-H, SIBTEST, IRT-LR, and LR, which are DIF determination techniques. According to Turkey - B-S-J-Z China comparison, all of the reading skills items that were accepted to show DIF showed a DIF effect in favor of China. Both Turkish and B-S-J-Z Chinese students took the adapted form of the tests. Accordingly, in the comparison of Turkey and B-S-J-Z China, which are the countries where the items adapted from the same source to different cultures were applied, it is noteworthy that all of the items showed DIF. The results of the Australia-Turkey comparison regarding whether the PISA 2018 Rapa Nui reading items show a DIF effect by culture variable are given in Table 9.

Table 9. The DIF Analysis Results of the Rapa Nui Unit Reading Skills Items of the PISA 2018 for Australia-Turkey Comparison

| | M-H Method | | SIBTEST | | IRT-LR | | Logistic | |
|---|---|---|---|---|---|---|---|---|
| | ΔMH | Level of effect | $X^2$ | Level of effect | $G^2$ | Level of effect | $X^2$ | Level of effect |
| M1 | 11.16 | A | 13.12 | B | 13.78 | B | 16.09 | A |
| M5 | 0.15 | - | 2.17 | - | 0.02 | - | 3.01 | - |
| M6 | 271.67 | C | 324.14 | C | 432.02 | C | 374.19 | C |
| M8 | 62.99 | C | 62.02 | C | 81.78 | C | 71.30 | A |
| M9 | 1.47 | - | 5.23 | A | 0.77 | - | 7.34 | A |
| M10 | 89.13 | C | 78.65 | C | 77.93 | C | 94.78 | A |
| M11 | 0.11 | - | 0.46 | - | 6.96 | A | 15.90 | A |

As seen in Table 9, items M1, M6, M8, and M10 showed a DIF effect in Australia and Turkey comparison in terms of M-H, SIBTEST, IRT-LR, and LR, which are DIF determination techniques. It was concluded that items coded M5, M9, and M11 did not show a DIF effect or showed a negligible level of DIF. Accordingly, when items coded M5 and M11 that did not show DIF were examined, it was seen that both were open-ended items. In the comparison of Australia and Turkey, all of the reading skills items accepted to show DIF showed a DIF effect in favor of Australia. The majority of items that showed DIF in all comparisons of PISA 2018 reading skills items were determined in Turkey and B-S-J-Z China comparison. Items that were found to show DIF in all comparisons were M6 and M8. However, a common item that did not show DIF in all comparisons was not determined. A bias analysis was conducted on items considered to show DIF for the Australia-Turkey comparison. Thus, it was aimed to determine whether the difference between cultures determined by DIF was a real difference or it was due to item bias.

### 3.1. Findings of the bias analysis of PISA 2018 Reading Skills Items that showed DIF in terms of culture variable

The bias analysis was carried out on the reading skills items of the PISA 2018 Rapa Nui unit which showed DIF according to the Australia-Turkey comparison. It was carried out in line with expert opinions. Accordingly, an expert opinion form, which allowed the experts to express their opinions on each item easily, was developed by the researcher. Experts evaluated each item to reveal whether they provided an advantage to one of the Australian or Turkish cultures, and, if they did, what the possible sources of bias were. Table 10 presents expert opinions about whether items provided an advantage to a cultural group and possible sources of bias.

Table 10. The distribution of expert opinions about whether the items that were accepted to show a DIF effect provided an advantage to a cultural group and possible sources of bias

| Expert opinions | | Items | | | | |
|---|---|---|---|---|---|---|
| | | M1 | M6 | M8 | M9 | M10 |
| | | f | f | f | f | f |
| It does not give an advantage to a cultural group. | | 7 | 5 | 6 | 7 | 6 |
| It gives an advantage to a cultural group. | | 10 | 12 | 11 | 10 | 11 |
| It gives an advantage to a cultural group | Australia | 10 | | 10 | 10 | 11 |
| | Turkey | - | 1 | 1 | - | - |
| Possible sources of bias | | | | | | |
| • The use of expressions or words in the item in different meanings | | 2 | 5 | 3 | 6 | 2 |
| • Familiarity of a culture group with the item content | | 10 | 6 | 3 | 4 | 4 |
| • Item format gives an advantage to a cultural group. | | 4 | 1 | 2 | - | 8 |
| • Cultural differences in skills measured by the item | | - | 3 | - | - | 1 |
| • Differentiation of items due to translation | | 4 | 7 | 2 | 3 | 2 |
| • Other | | 4 | 2 | 5 | 1 | 3 |

Table 10 contains the findings of the expert opinions taken to determine whether the PISA 2018 reading skills items showed item bias in terms of the culture variable. Accordingly, seven of the experts stated that the item coded M1 did not provide any advantage to any of the culture groups, while other seven stated that it provided advantage to Australian culture. Two experts indicated "the use of expressions or words in the item in different meanings" as a source of bias. For example, the expert coded U1 expressed opinion as follows: "*I think expressions such as 'which I learned to love', 'a good point for the beginning' may cause difficulties for Turkish students as the phrases contain direct translations.*" The expert coded U3 stated that Turkish students were at a disadvantage. "*The concept of blog may not be a very familiar word for that age group in Turkey. Mostly adults read blogs in Turkey, and the content of the blog is about places visited and food.*"

For the item coded M1, seven experts stated that "familiarity of a culture group with the item content" was a source of bias. Regarding this item, the expert coded U7 stated that the text provided advantage for Australian students.

> *"The regional details given in the introduction text can be time-consuming for Turkish students who are far from this area. Dealing with unnecessary details (where it is, its name, 3200 km, etc.) can cause a loss of meaning and time for them. In addition, it is observed that the island mentioned in the text (Easter Island- Australia) is a place that can be visited by tourists from Australia (this information can be accessed from the web)."*

According to the expert coded U11, "the word BLOG could have been written as a 'diary' or 'network diary' in Turkish." The expert coded U14 similarly stated that the word "blog" was disadvantageous to Turkish students. Arguing that the content of the item put Turkish students at a disadvantage, the expert coded U1 said, "The mention of Rapa Nui, Easter Island, and Moai exposed Turkish students to content that they were not familiar with." According to the expert coded U15, choosing a place like "Rapa Nui" was to the disadvantage of Turkish students. The expert coded U3 was also of the opinion that the names and places used in the text put Turkish students at a disadvantage. Four of the experts

stated that “the items differed due to translation. For example, the expert coded U7 highlighted important differences.

> “*It would be more appropriate in Turkish if it was translated as 'Professor's blog is presented on the side' instead of 'Use the professor's blog on the right'. I think there are similar instances where the translated form does not fit the meaning and flow of Turkish. A text that is obviously a translation is more difficult to read and understand holistically. In addition, the ‘23 May, 11.22’ statement on the blog is originally 11.22 a.m. The numbers ‘11.22’ separated by ‘.’ in the Turkish form actually tell the time. It may also be understood as December 2022 here. As a matter of fact, I was also confused at first. Since the first question asks time, we can state that this translation problem is confusing. It may cause the respondent to perceive the item as asking ‘how many months ago' according to the date given on the blog. Therefore, it can be problematic*.”

The expert coded U1 emphasized some translation problems saying, "*As I mentioned before, direct translation of some sentences without adaptation makes the meaning more difficult for Turkish students.*" The expert coded U10 stated that translation-bound disadvantages were observed in all items.

> “*There are sentence structures in the related text, not in the questions or options, which are frequently used in English but not so often in Turkish, and they sound a bit strange when used. For example, there are two noun clause structures in the first sentence. It is a common structure in English. In addition, since the subject and the predicate are located at the beginning of the sentence, it does not make the sentence as difficult to understand as it is in Turkish. The sentence has been translated into Turkish literally. I think that both the fact that such structures are not used frequently in Turkish and that they come between the subject and the predicate reduce the intelligibility of the sentence to the disadvantage of Turkish students. This was just an example. Word-by-word translation has been maintained throughout the entire text and similar disadvantages have emerged.*”

Experts who wanted to express an opinion different from the sources of bias stated in the expert opinion form expressed their opinions under the "other" option. The expert coded U2 made the following comments under this option:

> “*Since the Turkish translation is based on English, there are stylistic errors. For example, the expression of 'history class' is one of them. There is no such usage in Turkey or Turkish. This expression may have created an obstacle for Turkish students to understand the text quickly. Perhaps, it would be better if it was translated as 'you will attend this conference as a class as part of the history lesson.’ Another problematic statement was ‘it was written by the professor while he was living in Rapa Nui.’ The passive structure of the statement may have made it difficult to comprehend the sentence quickly.*”

Expert coded U11 highlighted a point very different from those of other experts. “*In the English version, the gender of the professor is clear with the use of the pronoun ‘she’ for females. There is no reference to the gender of the professor in the Turkish version. If gender has an effect in the reader's eyes in terms of a profession and carrying out that profession (it should be considered scientifically), it can be thought that this may also have a small effect.”* The opinion of the expert coded U12 regarding this item was as follows:

"*The expression "learn to love" in the English text appears as "which I learned to love" in the Turkish text, but this expression can be difficult to understand. In addition, the expression "civil war", whose culturally equivalent translation should be "iç savaş [iç: internal: savaş: war]", has been translated as "sivil savaş [sivil: civilian, savaş: war]" in the Turkish text. Cultural equivalence should be observed in the expressions used in the Turkish text to reduce such translation problems, and word-by-word translations should be avoided. Although the translations of some expressions are correct, it can be said that they look difficult in the Turkish form of the text because they have been translated word by word. Similarly, the term 'history class' is not similar to its equivalent in Turkish culture in terms of school structure, but only reflects the school system of English-speaking countries. In our case, it could have been translated as a history lesson or an excursion with the history teacher.*"

The examination of expert opinions for the item coded M6 indicated that 12 experts stated the item provided an advantage to a cultural group. Of these experts, only one stated it provided an advantage to Turkish culture, while 11 said it provided an advantage to Australian culture. Accordingly, four experts stated "the use of expressions or words in the item with different meanings" as a source of bias. The expert coded U5 stated that the item provided a disadvantage to Turkish culture as follows:

*"In the original form, the item is asking the student to identify facts and opinions in the text. This is a question that students can answer only if they know the meanings of these words. However, the concepts of objective and subjective have been used in the Turkish form. Therefore, the student needs to know what these concepts are to be able to answer the question. In other words, it does not simply make a distinction as to whether it is a real situation or an opinion, but a classification such as subjective or objective.*"

The expert coded U14 also stated that this situation may cause cross-cultural bias. The expert coded U8, on the other hand, stated that the item provided an advantage to Turkish students.

"*The expression 'gone' in the English expression 'the trees were gone' has been translated as "disappear" in the Turkish form, which is in favor of Turkish.*"

The expert coded U16 was of the opinion that the use of words with different meanings may have caused a bias.

"*I think the translation of the verb 'carve' into Turkish as "chip" makes it difficult for students to understand it. Using a more understandable expression such as 'carved and shaped' would help Turkish students to understand it better. They may not have understood that a Moi is a sculpture that is carved out of stone.*"

Pointing out that the familiarity of a cultural group with the item content is a source of bias, the expert coded U4 said, "*Whether the statement 'The book is written well and deserves to be read by anyone who is concerned about the environment' is objective or subjective can be understood by those who are familiar with English. The Turkish translation of the statement seems to be both subjective and objective and therefore it provides an advantage to the Australian culture.*"

The expert coded U10 stated that cultural differences in terms of skills measured by the item may be a source of bias and explained it as follows:

"*The subjectivity-objectivity may differ in terms of the cultural importance given to these concepts. While objectivity has a very distinct place in the Western understanding of education and science (I also include Australians since they migrated from Europe and largely preserve the same culture), the understanding of collectivism is more dominant in the Eastern culture, which also includes the Turkish society. Culturally, it doesn't make much difference whether a situation is subjective or objective. Subjective opinions are accepted as easily as scientific facts. But Western culture has struggled for objective understanding for centuries. The fact that concepts have different values will also affect the awareness of individuals about them.*"

The expert coded U8 made the following comments about the differentiation of items due to translation, thereby claiming the item provided an advantage for Turkish students.

"*In the original form, the adjective has been used after the noun in the expression 'the moai, the famous statues'. In the Turkish form, the adjective comes before the noun, making it easier to understand and saving time for the reader. In addition, the expression Moai has been written in capital letters and plural form in Turkish, while it has been written in lowercase and singular form in English. This provides an ease of understanding in favor of Turkish.*"

The expert coded U10 stated that the adapted form put Turkish students at a disadvantage as follows:

"*This is similar to the characteristics of items M1 and M5. A word-by-word translation has been done, and word choices or structures that are not very common in Turkish have been used in the translation of some of the sentences in the text. I think this may affect the intelligibility of the text.*"

The expert coded U1 made the following comments about the problems in the adapted form:

"*It may not be correct to say that some of the expressions in the items directly lead to differentiation. However, it would be better for Turkish students, if, for example, the statement 'it deserves to be read by everyone who is concerned about the environment' was translated as 'a book that everyone who is concerned about the environment should read.*'

The expert coded U2 stated the following opinion under the "other" option.

*"In this paragraph, there are expressions containing translation problems, although fewer than the previous ones. For example, the translation of 'about' in the first sentence is a word-by-word equivalent of the word. It could have been translated as "related to", which would meet the Turkish requirements more. In the second sentence, the verb 'what they did' must be singular because the subject of this verb, civilization, is a singular noun. The fluency that has been disrupted in the third sentence can be corrected with a connector after the word 'one'. Also, it could be 'Just after 700 AD' or 'early 700 AD' instead of 'some time after AD 700'; 'but the trees were gone' instead of 'but the trees disappeared'; 'hunted excessively' instead of 'overhunted', 'internal war' instead of 'civil war'."*

The expert coded U2 highlighted important points with the following comments about the text.

‘*The lesson to be learned from the book cannot be elicited from the statement 'The lesson to be learned from this wonderful but scary book is that in the past people preferred to destroy the environment by cutting down all trees and hunting animal species until they went extinct.' Instead, the sentence gives an idea about the topic of the book. On the other hand, the predicates connected by ‘and’ in the sentence ‘the book is well written and deserves to be read by everyone who is concerned about the environment’ are expected to be in the same mood. Predicates ending in “-ed” and “-ing” are not welcome in Turkish.*”

The expert coded U12 pointed out some important differences in terms of Turkish and English forms and the existence of some issues that may affect intelligibility.

“*The statement ‘The lesson to be learned from this wonderful but frightening book is that in the past people preferred to destroy the environment by cutting down all trees and hunting animal species until they went extinct.’ involves a noun clause structure connected by ‘that’. This can spoil the natural flow of Turkish texts. The best thing to do when translating these expressions is to split the sentence. By dividing the sentence, two smaller sentences are obtained. This usage is more common in our language. Combining two or three sentences that qualify each other will lead to semantic confusion. Such 'clause' structures are common in the English language and native English-speaking people are familiar with them. This provides an advantage to individuals whose native language is English. Similarly, the phrase, 'The book is well written and deserves to be read by anyone who is concerned about the environment,' seems to have different conjugations and tenses for the verbs 'written' and 'deserves'. This may make it difficult to understand the Turkish text.*”

The examination of the opinions of the experts on the item coded M8, which was accepted to show DIF in the comparison between Australia and Turkey, indicated that 10 experts stated that it provided an advantage for Australia and that one expert stated it provided an advantage in favor of Turkey. Six experts, on the other hand, stated that the item did not provide an advantage for any culture. Accordingly, the expert coded U1 mentioned the use of expressions or words in the item in different meanings as a source of bias as follows:

"*In fact, even the term 'Science reporter' is an expression that Turkish students are not familiar with.*"

The expert coded U5, on the other hand, mentioned the use of expressions or words in the item in different meanings as an advantage to Turkish students.

“*The expression ‘huge trees’ has been used in the English text, while the expression ‘large trees’ has been used in the options. While I was answering this question, I hesitated whether this meant something different. However, the Turkish form does not have such different expressions; both the text and the options read "giant trees.*"

The expert coded U13 also agreed on this opinion. The expert coded U1 mentioned the familiarity of a culture group with the item content as a source of bias.

"*I think that mentioning regions and names that are not common in Turkish culture may cause difficulties for students in terms of following the text*".

The expert coded U10 stated that the item differed due to translation and commented about the situation as follows:

"*It does not have the same meaning as in English. It's about the disappearance of trees. The verb 'disappear' has been used in the Turkish translation. The word-by-word translation of the reading text has led Turkish students to encounter expressions that they are not familiar with*".

The opinion of the expert coded U12 regarding the difference due to translation was as follows:

"*There are some expressions in the text that complicate the understanding due to translation, albeit partially. For example, there is a singular-plural mismatch in the translation of the following statement: 'Scientists agreed on the idea that giant trees disappeared when Europeans first arrived on the island in the 18th century but did not agree on Jared Diamond's theory about the cause of this extinction.' The translation of 'recently' could have been more accurate. To sum up, expressions must be translated more appropriately by paying attention to the grammatical structure and taking into account the principle of cultural equivalence, as I mentioned for the previous items.*"

Regarding the item coded M8, the expert coded U12 stated under the 'other' option that the politically correct translation of 'scientist' in Turkish did not refer to any gender.

"*It referred to male gender in the past, but it is genderless today. Translating it in the old way may cause a minor problem in understanding the text for younger generation students since this translation addresses the older generation. While reading the text, students can have a limited understanding by thinking only of male scientists. There may even be students who may react to this use.*"

According to the expert coded U2, who expressed opinion under the 'other' option, there were some translation problems. For example, translations about 'Diamond's theory about what happened in Rapa Nui', 'giant tree', and 'disappeared when they first arrived', could have been translated more appropriately. When the opinions of experts on the item coded M10, which was evaluated for bias, were examined, it was found that 11 out of 17 experts stated that the item provided an advantage for the Australian culture. Six experts thought that the item did not provide an advantage or a disadvantage for any culture. The expert coded U1, who stated that a cultural group's familiarity with the content of the item was a source of bias, said, "*As I mentioned in other parts, it would be easier to follow the text if the content was kept the same but the names of the scientists were changed to Turkish names. It can be much more difficult to follow the text due to foreign names."*

The expert coded U4 stated that the group taking the English version of the test had more advantages.

"*The English language group is more familiar with the information in the cause and effect box. Turkish speakers may have difficulty making some inferences and conclusions. It is as if the text does not provide enough clues*."

The expert coded U5 stated that the item format or stylistic features provided an advantage to a cultural group. The expert coded U3 claimed that there were cultural differences in terms of the skills measured by the item and that these skills were "Making inferences, establishing a cause-effect relationship, finding the supporting idea." The expert coded U10 stated that the items differed due to translation, the problem prevailed in other reading texts, and that the items had similar bias problems. The expert coded U12 emphasized that the problem in the

translation of the word ‘settlers’ put Turkish students at a disadvantage in terms of both the question and the options. “*The problem with the expression 'settlers' may apply to this question, as well. The translation of ‘settlers’ is missing in one of the boxes to be dragged into the blank in the root of the question.*”

Under the ‘other’ option, the expert coded U3 stated that there was a source of bias due to the item type and this was to the disadvantage of Turkish students.

“*Since students don't take enough computer-based tests in Turkey, they may have difficulty answering drag-and-drop type questions. Also, it measures high-level thinking skills. Students do not come across such items very often*.” The expert coded U12 stated that the translation of ‘scientist’ in most items could cause a disadvantage for Turkish students since the translation referred to the male gender.

## 4. Discussion and Conclusions

This part of the research included conclusions, discussions, and suggestions in line with the findings obtained within the scope of the research. This study was conducted to examine the cross-cultural measurement invariance of the items of the Rapa Nui unit in PISA 2018 Reading Skills test and the bias in the items that did not show invariance. As a result of the MGCFA conducted to examine the measurement invariance of the factor structure of the Rapa Nui unit on the samples of Australia, China, and Turkey, it was determined that the unit showed structural cross-cultural invariance. However, it was concluded that the Rapa Nui unit did not show metric invariance in terms of the three cultures. Analysis of metric invariance was repeated with paired combinations of Australian, Chinese, and Turkish cultures. Accordingly, it was concluded that the PISA 2018 reading skills Rapa Nui unit did not show cross-cultural metric invariance. In studies examining the cross-cultural measurement invariance of the booklets in the PISA 2012 mathematical literacy and PISA 2015 science literacy tests, it was concluded that the tests did not show metric invariance (Alatlı, 2020; Karakoç Alatlı, Çokluk Bökeoğlu, 2018; Oliden and Lizaso (2013) concluded that the PISA 2009 reading skills test showed metric invariance according to different languages. Söyler (2020) determined that the PISA 2015 reading skills test did not show measurement invariance in terms of native and non-English speaking countries. Elosua and Mujika (2013) concluded that the PISA 2009 reading skills test showed metric invariance for different languages in the same country.

In this study, the DIF status of seven items in the "Rapa Nui" unit of the PISA 2018 reading skills test was examined in terms of different cultures. For this purpose, the results related to Australia - B-S-J-Z China, Turkey - B-S-J-Z China, and Australia-Turkey comparisons respectively by using M-H, SIBTEST, IRT-LR, and Logistic Regression, which are DIF determination techniques, were discussed. In the comparison of Australian and B-S-J-Z Chinese cultures regarding the seven items in the "Rapa Nui" unit in the PISA 2018 reading skills test, five of the seven items (M5, M6, M8, M9 and M11) were determined to show DIF in favor of China. The items showing DIF had no similarity in terms of cognitive processes and item content. Items coded M1 and M10, which did not show DIF, were multiple-choice questions. According to the Turkey - B-S-J-Z China comparison, it was determined that all seven items examined within the scope of the research showed DIF in favor of China. In the comparison of Australia - B-S-J-Z China and Turkey-China made in terms of two cultures that took the adapted form of the items, it is noteworthy that the items were in favor of the Chinese culture, which had a higher PISA 2018 reading skills

performance. In the comparison of Australia-Turkey, four out of seven items (M1, M6, M8, and M10) were found to show DIF. All of these items, which were found to show DIF, were multiple-choice questions. On the other hand, items coded M5 and M11 that did not show DIF in the comparison between Australia and Turkey were open-ended questions. In the study by Özmen (2014), in which PISA 2009 reading skills items were examined in terms of item bias according to Turkey, the USA, and the UK samples, it was determined that approximately 40% of the items showed DIF. Ceyhan (2019) examined the measurement invariance of the PISA 2012 reading skills test in terms of the same language, different culture, and different language and different cultures. Accordingly, it was determined that structural invariance was provided for the same languages, and a weak invariance was provided for different languages. In the study conducted for the PISA 2000 reading skills items, fewer items were found to show DIF in the same language and different countries comparisons compared to the same country and different language comparisons (Grisay & Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009). This shows that language is an important variable for DIF. In this study, some items were found to show DIF as a result of comparisons of different cultures and different languages.

An item bias analysis was conducted on the items determined by DIF determination techniques and accepted as showing DIF in the comparison between Australia and Turkey. The results of the expert opinions on the item bias analysis were presented in the study. According to the 17 experts who expressed their opinions within the scope of the research, it was concluded that all four items accepted as showing DIF among the items of PISA 2018 reading skills Rapa Nui unit test showed a bias in favor of Australia. When the sources of bias regarding the items were examined, it was determined that they were “familiarity of a culture group with the item content”, “differentiation of items due to translation”, “the use of expressions or words in the item in different meanings”, “other”, and “item format gives an advantage to a cultural group” in the descending order. Özen (2013) determined cultural differences, familiarity with the item format, difficulty understanding the words, and translation errors as the sources of bias for the items of the PISA 2009 reading skills test based on expert opinions. Grisay and Monseur (2007) concluded that asking more than one question based on the same reading text caused local dependence and this, in turn, resulted in item bias in PISA 2000 and 2001 reading skills items. In addition, they also stated that translation was also a source of bias. Asil and Brown (2016) examined the measurement invariance of the PISA reading skills test in terms of 55 countries, including Australia as the reference group. Accordingly, they determined that the socio-economic source of education played an important role in the measurement invariance. They added that language factors and educational practice played a smaller role in measurement invariance.

## 5. Recommendations

In line with the results obtained from the research, suggestions for practitioners were discussed primarily. The results obtained from large-scale applications such as PISA, TIMSS, and PIRLS are effective in education reforms of countries and cross-cultural comparisons. For this reason, that the tests and item groups used in these applications show measurement invariance according to many variables is highly important, considering the areas in which the results are used. For this purpose, cross-cultural invariance is an important measurement invariance assumption for adaptation studies. Therefore, item writing, pilot implementation, adaptation, and similar processes must be carried out painstakingly. Invariance analyses should be carried out on pilot application data and changes should be made in line with the sources of bias. In particular, in line with the results obtained in this study, such variables

should be taken into account in the item writing process so that the familiarity of a culture group with the item content will not be seen as a source of bias. Translation problems were identified as an important source of bias in all items. For this reason, the adaptation process should be carried out meticulously, especially in the target culture. At this point, care should be taken in terms of possible sources of item bias in both item writing and adaptation processes, and if necessary, education should be given for this purpose.

In line with the results obtained from this research, it is primarily recommended to researchers that similar studies should be repeated in different cultures and with different DIF techniques. However, since the individualized test application was used in PISA 2018, there were no students taking the same test, so the analyses could be carried out on the items. For this reason, DIF and item bias studies should be carried out in terms of different item clusters, as well.

## References


Acar, T. (2008). *Maddenin Farklı Fonksiyonlaşmasını Belirlemede Kullanılan Genelleştirilmiş Aşamalı Doğrusal Modelleme, Lojistik Regresyon ve Olabilirlik Oranı Tekniklerinin Karşılaştırılması.* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi/Eğitim Bilimleri Enstitüsü, Ankara.

Alatlı, B. (2020). Cross-cultural Measurement Invariance of the Items in the Science Literacy Test in the Programme for International Student Assessment (PISA-2015*). International Journal of Education & Literacy Studies, 8*(2). doi: doi:10.1023/A:1005139318357

Arffman, I. (2010). Equivalence of Translations in İnternational Reading Literacy Studies. *Scandinavian Journal of Educational Research, 54(1),* 37-59.

Asil, M. & Brown G., T., L. (2016) Comparing OECD PISA Reading in English to Other Languages: Identifying Potential Sources of Non-Invariance, *International Journal of Testing, 16*(1), 71-93, doi: 10.1023/A:1005139318357

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.

Brown, T. A. (2006). *Confirmatory Factor Analysis For Applied Research*. New York: The Guilford Press.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.

Ceyhan, E. (2019). *Pısa 2012 Okuma Becerileri Ölçeğinin, Uygulama Dili Doğrultusunda Belirlenen Ülkeler Arasında Ölçme Değişmezliğinin İncelenmesi*. Akdeniz Üniversitesi/ Eğitim Bilimleri Enstitüsü, Antalya.

Cheung, G. W. & Rensvold, R. B. (2000). Assessing Extreme And Acquiescence Response Sets İn Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-cultural Psychology, 31*(2), 188-213. doi: 10.1177/0022022100031002003

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling, 9,* 233–255. doi:10.1207/S15328007SEM0902_5

Çepni, Z. (2011*). Degişen Madde Fonksiyonlarının SIBTEST, Mantel Haenszel, Lojistik Regresyon ve Madde Tepki Kuramı yöntemleriyle İncelenmesi.* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi/Eğitim Bilimleri Enstitüsü, Ankara.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk , Ş. (2010*). Sosyal Bilimler için Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları*. Ankara: Pegem Akademi.

Doğan, N., & Öğretmen, T. (2008). Degişen Madde Fonksiyonunu Belirlemede Mantel - Haenszel, Ki-kare ve Lojistik Regresyon Tekniklerinin Karşılaştırılması. *Egitim ve Bilim, 33,* 100-112.

Dorans, N. J. & Holland, P. W. (1993). DIF Detection And Description: Mantel-Haenszel and Standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). New Jersey: Lawrence Erlbaum Associates Publishers.

Elosua, O. P., & Mujika, L. J. (2013). Invariance Levels Across Language Versions of the PISA 2009 Reading Comprehension Tests in Spain. *Psicothema, 25 (3),* 390–395.

Elosua, P., & López-Jaúregui, A. (2007). Potential Sources of Differential Item Functioning in the Adaptation of Tests. *International Journal of Testing, 7*(1), 39–52. doi:10.1080/15305050709336857

Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research, 42,* 509-529. doi: 10.1080/00273170701382864

Fan, X., Thompson, B., & Wang, L. (1999). Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. *Structural Equation Modeling*, *6*(1), 56–83. doi: 10.1080/10705519909540119

Fraenkel, J.R., & Wallen, N.E. (2006). *How to Design and Evaluate Research in Education.* 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. (2000). Construct Equivalence on Translated Achievement Tests. *Canadian Journal of Education, 25*(4), 280-296. doi: 10.2307/1585851

Gotzmann, A., Wright, K., & Rodden, L. (2006, April). *A Comparison Of Power Rates For İtems Favoring The Reference And Focal Group For The Mantel-Haenszel And Sıbtest Procedures.* Paper presented at the American Educational Research Association (AERA) in San Francisco, California.

Gök, B., Kelecioglu, H. & Doğan, N. (2010). Değişen Madde Fonksiyonunu Belirlemede Mantel-Haenzsel ve Lojistik Regresyon Tekniklerinin Karşılaştırılması. *Eğitim ve Bilim, 35,* 3-16.

Greer, T. G. (2004). *Detection of differential item functioning (dif) on the satv: A comparison of four methods: Mantel-Haenszel, Logistic Regression, simultaneous item bias and likelihood ratio test* (Unpublished doctoral dissertation). University of Houston

Grisay, A. & Monseur, C. (2007). Measuring the Equivalence of Item Difficulty in the Various Versions of an International Test. *Studies in Educational Evaluation, 33,* 69-86. http://www.sciencedirect.com/science/article/pii/S0191491X07000077?#

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of Item Difficulties Across National Versions of the PIRLS and PISA Reading Assessments. *ERI Monograph Series: Issues and Methodologies In Large-Scale Asssesments, 2,* 63-84.

Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. *Medical Care, 44,* 182-188.

Hambleton, R.K. & Swaminathan, H. (1989). Item Response Teory: Principles And Applications. USA: Kluwer Nijhoff Publishing

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D., (2005). *Adapting Educational and Psychological Tests for Cross-cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R.K., Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

He, J., & Van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. *Online Readings in Psychology and Culture, 2(2).* doi: 10.9707/2307-0919.1111

Holland, P. W., & Thayer, D. T. (1988). Differentialitem performance and the Mantel-Haenszel proce-dure. In H. Wainer & H. I. Braun (Eds.), Test va-lidity (pp. 129-145). Hillsdale NJ: Erlbaum.

Hu, L., & Bentler, P. (1995). Evaluating Model Fit. R. Hoyle (Ed). *Structural Equation Modeling: Concepst, Issues and Application* (pp. 76-99). Thousand Oasks: Sage Publications.

Jodoin, M. G. & Gierl, M.J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure with Logistic Regression Procedure Or DIF Detection. *Applied Measurement in Education, 14*(4), 329-349. doi:10.1207/S15324818AME1404_2

Johnson, T. P. (1998). Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research. *ZUMA-Nachrichten Spezial, 3,* 1-40. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49730-6

Karakoç-Alatlı, B. ve Çokluk-Bökeoğlu, Ö. (2018). Investigation of measurement invariance of literacy tests in the Programme for International Student Assessment (PISA-2012). *Elementary Education Online, 17*(2), 1096-1115. doi: 10.17051/ilkonline.2018.419357

Karasar, N (2011). *Bilimsel Araştırma Yöntemi*. Ankara: Nobel Yayın Dağıtım

Kreiner, S., & Christensen, K. B. (2014) Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika, 79(2),* 210—231

Lei, M., & Lomax, R. G. (2005). The Effect of Varying Degrees of Nonnormality in Structural Equation Modeling. *Structural Equation Modeling, 12*(1), 1–27. doi:10.1207/s15328007sem1201_1

Little, T.D. (1997). Mean and Covariance Structures (MACS) Analyses of Crosscultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Reseach. 32,* 53- 76.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Magis, D., Béland, S., Tuerlinckx, F. & *Boeck P., D.* (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42,* 847–862. doi:10.3758/BRM.42.3.847

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/The Faculty of Graduate Studies, Vancouver.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results*. Paris: OECD Education Working Papers (EDU/PISA/GB(2008)28).

MEB (2019). PISA 2018 ulusal ön raporu. Ankara http://pisa.meb.gov.tr/www/pisa-2018-turkiye-on-raporu-yayimlandi/icerik/3

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297–334. doi:10.1177/014662169301700401

OECD (2019). PISA 2018 technical report. Paris: OECD Publications. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

Oliden, P. E. & Lizaso, J, M. (2013) Invariance Levels Across Language Versions of the PISA 2009 Reading Comprehension Tests in Spain. *Psicothema, 25(3),* 390-395 doi: 10.7334/psicothema2013.46

Oliveri, M. E. & von Davier, M. (2011). Investigation of Model Fit and Score Scale Comparability in International Assessments. *Psychological Test and Assessment Modeling, 53(3),* 315–333.

Osterlind, S. J. & Everson, H., T. (2009). *Differential Item Functioning. (2nd).* Thousand Oaks. CA: SAGE Publications

Özmen, D. T. (2014). PISA 2009 Okuma Testi Maddelerinin Yanlılığı Üzerine Bir Çalışma. *Eğitim Bilimleri ve Uygulama, 13 (26),* 147-165.

Prelow, H. , Tien, 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*, 225-244. doi:10.1023/A:1005139318357

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.

Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Equivalence. *Psychological Bulletin, 114,* 552-566.

Rizopoulos, D. (2006). ltm: An R Package For Latent Variable Modelling and İtem Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1–25. Doi: 10.18637/jss.v017.i05

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159–194. doi:10.1007/BF02294572

Sireci, S. G. & Swaminathan, H. (1996). *Evaluating translation equivalence: so what's the big dif*? Paper Presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY.

Söyler Bağdu, P. (2020). *PISA 2015 Okuma Becerileri Testinin Ana Dili Değişkenine Göre Ölçme Değişmezliğinin İncelenmesi.* Ege Üniversitesi/Eğitim Bilimleri Enstitüsü, İzmir.

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting Differential Item Functioning with Comfirmatory Factor Analysis and Item Response Theory: Toward A Unified Strategy. *Journal of Applied Psychology, 91(6),* 1292- 1306.

Stout, W. & Roussos, L. (1995). *SIBTEST User Manual*. Urbana: University of Illinois.

Swaminathan, H. & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*(4), 361–370. doi:10.1111/j.1745-3984.1990.tb00754.x

Şencan, H. (2005). *Sosyal ve Davranışsal Ölçümlerde Güvenirlik ve Geçerlilik.* Ankara: Seçkin Yayınları

Şimşek, Ö. F. (2007). *Yapısal Eşitlik Modellemesine Giriş: Temel ilkeler ve LISREL Uygulamaları*. Ankara: Ekinoks

Tabachnick, B. G. & Fidell, L. S. (2007). Using Multivariate Statistics. Fifth Edition. Pearson: AB

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.

Walker, M. (2007). Ameliorating Culturally-Based Extreme Response Tendencies to Attitude Items. *Journal of Applied Measurement, 8,* 267-278.

Wetzel, E. & Carstensen, C. H. (2013). *Linking PISA 2000 and PISA 2009: Implications of Instrument Design on Measurement Invariance. Psychological Test and Assessment Modeling, 55(2),* 181-206.

Wu, A. D., Li, Z. & Zumbo, B. D. (2007). Decoding The Meaning of Factorial Invariance and Updating The Practice of Multigroup Confirmatory Factor Analysis: A Demonstration with TIMSS Data. P*ractical Assessment, Research and Evaluation, 12(3),* 1-26. doi: 10.7275/mhqa-cd89

Zumbo, B. D. & Thomas, D. R. (1996). *A Measure of DIF Effect Size Using Logistic Regression Procedures.* Paper presented at National Board of Medical Examiners, Philadelphia, PA.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework Forbinary and Likert-type (ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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