2018, 12(2), 152- 172
Novitas-ROYAL (Research on Youth and Language)

The Examination of Measurement Invariance and Differential Item Functioning of PISA 2015 Cognitive Tests In Terms of The Commonly Used Languages

Güçlü ŞEKERCİOĞLU[1] & Hakan KOĞAR[2]

[1] Ph.D., Department of Educational Sciences, Akdeniz University, Turkey, guclus@akdeniz.edu.tr
[2] Ph.D., Department of Educational Sciences, Akdeniz University, Turkey, hkogar@gmail.com

Article information

Submission	15/02/2018
Revision received	25/05/2018
Acceptance	11/08/2018

Keywords
PISA, large-scale tests, measurement invariance, differential item functioning, language

Abstract: The aim of the present study was to examine the measurement invariance (MI) of the reading, mathematics, and science tests in terms of the commonly used languages. It also aimed to examine the differential item functioning (DIF) of the PISA test, the original items of which are in the languages of English and French, in terms of the language variable. As a result of the MI analyses, it was concluded that the best working model for the reading comprehension, mathematics and science tests in terms of the commonly used languages for all three tests was the strong factorial invariance model. As a result, it was found that MI could not be established for any of these tests. Moreover, the analyses indicated that population heterogeneity (PH) could not be obtained in any of the three tests, and that the latent means for these tests were not equal. Furthermore, DIF was identified in more than half of the items. It can be deduced that this condition may result in an increase in the number of items with DIF specifically in terms of the variety in the number of languages.

Anahtar sözcükler
PISA, geniş ölçekli sınavlar, ölçme değişmezliği, değişen madde fonksiyonu, dil

Yaygın Kullanılan Diller Açısından PISA 2015 Bilişsel Testleri İçin Ölçme Değişmezliğinin ve Değişen Madde Fonksiyonlarının İncelenmesi
Öz: Bu araştırmanın amacı, PISA 2015 uygulamasına ait okuduğunu anlama becerisi, matematik ve fen okuryazarlığı testlerinin yaygın kullanılan diller açısından ölçme değişmezliğini ve söz konusu testlere ait maddeler için orijinal dili olan İngilizce ve Fransızca olan PISA'nın dil değişkenine göre değişen madde fonksiyonunu incelemektir. Bu doğrultuda PISA uygulamasında orijinal diller dışında yaygın kullanılan beş test dili belirlenmiştir. Ölçme değişmezliği analizleri sonucunda okuduğunu anlama becerisi, matematik ve fen okuryazarlığı testlerinin üçü için de yaygın kullanılan diller açısından en iyi çalışan modelin güçlü faktöriyel değişmezlik modeli olduğu ve dolayısıyla bu testler için ölçme değişmezliğinin sağlanamadığı sonucuna ulaşılmıştır. Diğer taraftan maddelerin yarısından çoğunda DMF tespit edilmiştir. Bu durum, özellikle dil sayısındaki çeşitlilik ile birlikte değişen madde fonksiyonuna sahip madde sayısının da artmasına neden olabileceği biçiminde düşünülebilir.

1. Introduction

The standardized large-scale international assessments play an important role, especially for decision makers and policy makers. Large-scale tests are not only used to identify the strengths and weaknesses of the education system of a certain country but also to compare different countries' education systems. Many countries allocate a large amount of funds to these large-scale tests and place a high level of confidence on the results of these tests. For this reason, the findings obtained from large-scale tests are very important.

Some important problems in making comparisons among countries, which is a key feature of large-scale tests, are discussed in the related literature (Huang, 2010; Huang, Wilson, & Wang, 2016; Roth, Ercikan, Simon, & Fola, 2015). To ensure the validity of the comparison of scores obtained from these large-scale tests, the participant groups (or subgroups) need to have the same or a similar level of knowledge and skills. From this perspective, the Programme for International Student Assessment (PISA), which is one of the large-scale studies, conduct their research on 15-year-old students. However, this condition is not sufficient for making valid comparisons. The participant groups should be comparable not only in the measured trait, but also in some aspects of assessment such as bias.

Bias is the condition where the test scores obtained from the groups with various traits (gender, culture, language, etc.) are not equivalent (Van de Vijver & Poortinga, 1997). On the other hand, item bias emerges in relation to the content of the items or is based on communication. In educational research studies, students who participate in different education programs may not be familiar with certain topics. When students speaking different languages are assessed, bias may occur due to translation errors (Marotta, Tramonte, & Willms, 2015). The increase in the difficulty of understanding words chosen within the course of a bad translation process particularly challenges young students. Hence, difficulties in understanding words negatively impacts interpretation as well (He & Wolfe, 2010).

The most probable source of bias in the scores obtained from large-scale tests is attributed to multilingualism and multiculturalism (Huang, Wilson, & Wang, 2016). The equality of the test is jeopardized by the differentiation in students' familiarity with the content and the topic of the questions derived from translation or errors in translation. Moreover, it is jeopardized by means of conditions in which some items are found difficult or easy by different students having the same level of ability in different countries. However, by conforming to the rules of international studies and with the support of participant countries, the exact equality of different language versions of tests can be ensured. Despite these measures, numerous studies report that the comparability of the scores obtained from the different language versions of tests is under risk (Allalouf, Hambleton, & Sireci, 1999; Grisaya & Monseur, 2007; Price & Oshima, 1998). This condition shows that it is essential to establish measurement invariance (MI) so that valid comparisons can be made among students who are assessed in different languages. By MI, synchronous assessment of the equality of structural parametres for more than one group can be obtained. In the related literature, it is possible to encounter different terms for different tests of MI, such as the *factor structure equality test*, *metric invariance* and *factorial invariance*. In addition, the examination of the structural congruence also leads to the assessment of the concept of population heterogeneity (Brown, 2006).

Numerous studies have reported problems in the accuracy and validity of the comparability of test scores deriving from the language of the test (Abedi, 2004; Abedi, Hofstetter, & Lord, 2004; Butler, Bailey, Stevens, Huang, & Lord, 2004; Nguyen & Cortes, 2013). There are also various tests reporting that MI cannot be established in different languages and different

countries (Allalouf, 2000; Ercikan & Koh, 2005; Price & Oshima, 1998; Xie & Wilson, 2008). However, there are few studies that have conducted differential item functioning (DIF) analyses in order to identify potential item biases in conditions where MI is not established (Alatlı & Bökeoğlu, 2018; Ardic & Gelbal, 2017; French, Finch, & Vazquez, 2016; Huang, Wilson, & Wang, 2016; Oliveri, Ercikan, & Zumbo, 2013). Furthermore, the number of studies that have focused on the language and translation of the test is quite limited (Alatlı & Bökeoğlu, 2018; Huang, Wilson, & Wang, 2016; Oliveri, Ercikan, & Zumbo, 2013).

Alatlı and Bökeoğlu (2018) have examined the MI of the reading comprehension, mathematics and science literacy subtests in PISA 2012 with respect to the language variable. It was found that structural invariance was established in the tests, while metric invariance was not. In the math literacy, science literacy and reading comprehension tests, the items with DIF in terms of the language variable were found to constitute 35%, 34% and 22% of the test, respectively. The tests including the highest degree of DIF were the English and Chinese versions. In a study aiming to identify these sources of DIF deriving from translation, Huang, Wilson, and Wang (2016) identified four fundamental language-based sources of DIF. These are the insufficiencies in students' vocabulary in understanding the key word choices, the length of the texts, the grammatical structures of key sentences, and the contextual meanings of the texts. In the study they conducted with the data set of PIRLS 2006, Oliveri, Ercikan, and Zumbo (2013) examined the DIF source with respect to the Chinese and Arabic languages. The students were distributed heterogeneously based on the language they spoke into three different latent classes determined according to latent class analyses.

The aim of the present study was to examine the MI of the selected cognitive test forms of reading comprehension, mathematical literacy, and science literacy tests of PISA 2015 in relation to the most commonly used languages (English, Spanish, Arabic, Portuguese, Chinese, German, French, and Turkish). It also aimed to examine the differential item functioning of the items in the same tests with respect to the language variable (the original languages of English and French).

2. Method

2.1. Data

In the present study, the PISA 2015 data set, provided to the researchers by OECD, were used. Seven commonly used test languages (minimum 3% and above) were identified in PISA, which is implemented in 47 languages with the participation of a total of 519,334 students. These are English, Spanish, Arabic, Portuguese, Chinese, German and French. These test languages constitute 59.14% of the data in total (n=307124). When Turkish is also added to this data, the ratio increases to 60.28% (n=313080). The data of English, which is the most commonly used language, constitute 17.52% of the total data (n=90996). The countries where the language of the test is English are Australia, Canada, Cyprus, Hong Kong, Luxembourg, Macau, Malaysia, New Zealand, Scotland, Singapore, Sweden, the United Arab Emirates, the United Kingdom, and the Unites States of America. The data of Spanish, which is the second most commonly used language, constitute 16.33% of the total data (n=84797). The countries where the language of the test is Spanish are Chile, Columbia, Costa Rica, Mexico, Peru, Spain, and Uruguay. The data of Arabic, which is the third most commonly used language, constitute 6.6% of the total data (n=34283). The countries where the language of the test is Arabic are Israel, Tunisia, and the United Arab Emirates. The data of Portuguese, which is the fourth most commonly used language, constitute 5.65% of the total data (n=29348). The countries where the language of the test is Portuguese are Brazil, Macau, and Portugal. The data of Chinese, which is the fifth most commonly used language, constitute

5.07% of the total data (n=26354). The places where the language of the test is Chinese are China, China Taipei, and Macau. The data of German, which is the sixth most commonly used language, constitute 4.23% of the total data (n=21946). The countries where the language of the test is German are Austria, Belgium, Germany, Italy, and Luxembourg. Finally, the data of French, which is the seventh most commonly used language, constitute 3.74% of the total data (n=19400). The countries where the language of the test is French are Canada, Belgium, France, and Luxembourg. On the other hand, the tests in Turkish are implemented only in Turkey, and the Turkish data constitute 1.15% (n=5956) of the total data.

In the analyses of the reading comprehension scores, the data were obtained from booklet number 41, and in the analyses of mathematical literacy scores, the data were obtained from booklet number 43. The data sets used in the analyses of reading comprehension and mathematical literacy are relatively simple, so they were easily selected from a single booklet for each analysis; however, the data sets for the analysis of science literacy scores were more complex, so six data sets were obtained, of which the 5th data set was selected for use in this study. In the data set selected for analyses of science literacy scores, the data were obtained from booklet numbers 33, 42, 45, 54 and 93. As the number of participants taking the tests in Turkish was low, the data set in which the number of participants in the Turkish test were high were taken into consideration for comparability in the selection of the Turkish booklets.

With respect to the reading comprehension test of PISA 2015, analyses were conducted with the data of 7,265 participants, in accordance with the study's primary and secondary aims. Of the participants, 29.9% had taken the test in English (n=2175), 24.3% in Spanish (n=1769), 9.3% in Arabic (n=675), 11.1% in Portuguese (n=807), 9.6% in Chinese (n=694), 8% in German (n=583), 5.5% in French (n=403) and 2.2% in Turkish (n=159).

With respect to the mathematical literacy test, analyses were conducted with the data of 7,855 participants. Of the participants, 29.4% had taken the test in English (n=2308), 28% in Spanish (n=2199), 8.3% in Arabic (n=652), 9% in Portuguese (n=706), 9.2% in Chinese (n=719), 8% in German (n=626), 6.1% in French (n=477) and 2.1% in Turkish (n=168).

Finally, with respect to the science literacy test, analyses were conducted with the data of 7,008 participants. Of the participants, 28.2% had taken the test in English (n=1979), 27.9% in Spanish (n=1954), 7.9% in Arabic (n=552), 11% in Portuguese (n=772), 9.6% in Chinese (n=672), 7.5% in German (n=529), 5.9% in French (n=413) and 2% in Turkish (n=137).

2.2. Data Analysis

In accordance with the aim of the study, descriptive statistics, KR-20 reliability analysis and an equality test of covariance matrices were performed to test the statistical assumptions. Confirmatory factor analysis (CFA) and multi-group CFA were performed to seek the responses to the research questions regarding MI, while Mantel-Haenszel's methods and Raju's two-parameter logistic (2PL) models were employed to seek responses to the research questions regarding DIF.

Multi-group CFA was conducted to determine whether or not MI was established in the reading comprehension, mathematical literacy and science literacy test scores. Prior to the MI analyses, the test statistics, normality tests, and reliability coefficients of the groups under study were initially calculated in accordance with the basic assumptions of the analysis.

Table 1
Test Statistics, Normality Tests, and Reliability Coefficients for the PISA Reading Comprehension Test Scores

Form Lang.	*n*	*Mean*	*Median*	*Mode(s)*	*SD*	*Range*	*Skw.*[1]	*Kur.*[2]	*KR-20*
English	2175	15.76	16	18	6.06	28	-.28	-.71	.87
Spanish	1769	13.21	13	12	5.87	27	-.03	-.76	.86
Arabic	675	9.36	9	4 & 7	4.89	24	.31	-.52	.91
Portuguese	807	10.32	10	10	5.87	27	.36	-.56	.93
Chinese	694	16.51	17.5	19	5.38	27	-.65	-.11	.85
German	583	15.69	16	17	5.77	26	-.40	-.57	.86
French	403	15.88	17	21	5.97	27	-.47	-.50	.87
Turkish	159	11.65	12	11	4.97	22	.12	-.68	.82

[1] Skewness [2] Kurtosis

As can be seen in Table 1, it can be maintained that in terms of total scores, the measures of central tendency for the groups who took the PISA reading comprehension test in different languages are relatively close to each other. The coefficients of skewness and kurtosis are within ∓ 1 interval which indicates that the distribution is close to normal (Rosenthal & Rosnow, 2008). Even though the coefficients presented in Table 1 are within ∓ 1 interval, it can be stated that the distribution of the Arabic, Portuguese and Turkish test scores are somewhat positively skewed, while the distribution of the scores of the other languages are somewhat negatively skewed. As for the internal reliability coefficients, calculated based on the scores obtained from the tests of the groups, it can be observed that they are generally at an acceptable level. According to Nunnaly and Bernstein (1994), a reliability coefficient between .70 and .80 is adequate for studies to be considered reliable. In the present study, the KR-20 reliability coefficients for all the tests were found to be above .80.

Moreover, the equality of covariance matrices, one of the multi-group CFA assumptions, was tested once again for the reading comprehension test. As a result of analysis, the following data set was found: S-B$\chi^2(2842)=8888.57$, S-B$\chi^2/df=3.13$, RMSEA=.048, CFI=.94, and SRMR=.11. Accordingly, the degree of freedom and S-Bχ^2 are around 3, RMSEA is below .05, CFI is above .90 and SRMR is above .08. A general examination of the fit indices indicates that there is a fit among the eight-covariance matrices.

Based on the basic assumptions of the analysis, the values obtained from the test statistics, normality tests, and reliability coefficients of the mathematical literacy test of the groups under study are presented in Table 2.

Table 2
Test Statistics, Normality Tests, and Reliability Coefficients for the PISA Mathematical Literacy Test Scores

Form Lang.	*N*	*Mean*	*Median*	*Mode*	*SD*	*Range*	*Skw.*	*Kur.*	*KR-20*
English	2308	9.43	9	9	4.55	22	.28	-.63	.82
Spanish	2199	6.86	6	4	4.18	21	.70	-.16	.81
Arabic	652	4.15	3	3	2.99	17	1.20	1.53	.71
Portuguese	706	5.94	5	2	3.92	20	.92	.52	.80
Chinese	719	12.14	13	13	4.84	21	-.15	-.85	.84
German	626	10.13	10	10	4.66	21	.05	-.79	.83
French	477	9.60	9	6	4.59	21	.13	-.82	.84
Turkish	168	6.04	5	3	3.88	18	.93	.50	.78

As can be seen in Table 2, it can be maintained that in terms of total scores, the measures of central tendency for the groups who took the PISA mathematical literacy test in different languages are relatively close to each other. Even though the skewness and kurtosis coefficients are within $\mp$1 interval for all the commonly used languages except Arabic, it can be stated that the distribution of the Chinese test scores are somewhat negatively skewed, while the distribution of the scores of the other languages are somewhat positively skewed.
The distribution for the Arabic data can be said to be positively skewed, peaked, and exceeding the tolerance limit. As for the internal reliability coefficients, calculated based on the scores obtained from the tests of the groups, it can be observed that they are generally at an acceptable level. In other words, the KR-20 reliability coefficients for all the tests, except for the Arabic and Turkish mathematical literacy tests, were found to be above .80.

The equality of the covariance matrices for the mathematical literacy test was tested. As a result of analysis, the following was found: S-Bχ^2(1771)=14753.3, S-Bχ^2/df=8.33, RMSEA=.086, CFI=.79, and SRMR=.11. Accordingly, the degree of freedom with S-Bχ^2 is around 5, RMSEA is above .08, CFI is below .90 and SRMR is above .08. An overall examination of the fit indices indicates that there is no fit among the eight-covariance matrices.

Finally, based on the basic assumptions of the analysis, the values obtained from the test statistics, normality tests and reliability coefficients of the science literacy test for the groups under study are presented in Table 3.

Table 3
Test Statistics, Normality Tests, and Reliability Coefficients for the PISA Science Literacy Test Scores

Form Lang.	*N*	*Mean*	*Median*	*Mode(s)*	*SD*	*Range*	*Skw.*	*Kur.*	*KR-20*
English	1979	16.47	17	17 & 23	7.33	33	-.06	-.94	.89
Spanish	1954	12.25	11	8	6.80	31	.43	-.63	.88
Arabic	552	8.84	8	5	5.66	29	.85	.44	.87
Portuguese	772	9.21	8	5	6.52	32	.88	.19	.88
Chinese	672	19.14	20	27	6.74	33	-.44	-.53	.86
German	529	16.30	16	13	6.75	29	.03	-.92	.86
French	413	16.15	16	19 & 22	7.13	32	-.11	-.96	.88
Turkish	137	10.70	10	11	5.43	26	.63	-.14	.82

As can be seen in Table 3, it can be maintained that in terms of the total scores, the measures of central tendency for the groups who took the PISA science test in different languages are relatively close to each other. Even though the skewness and kurtosis coefficients are within $\mp$1 interval for all of the commonly used languages, it can be stated that the distribution of the English, Chinese, and French test scores are somewhat negatively skewed, while the distribution of the scores of the other languages are somewhat positively skewed. As for the internal reliability coefficients, calculated based on the scores obtained from the tests of the groups, it can be observed that they are generally at an acceptable level. In other words, the KR-20 reliability coefficients for all the tests were found to be above .80.

The equality of the covariance matrices for the science literacy test was tested. As a result of analysis, the following was found: S-Bχ^2(4165)=10608.53, S-Bχ^2/df=2.55, RMSEA=.042, CFI=.95, and SRMR=.089. Accordingly, the degree of freedom with S-Bχ^2 is below 3, RMSEA is below .05, CFI is equal to .95 and SRMR is above .08. An overall examination of the fit indices indicates that there is fit among the eight-covariance matrices.

MI was tested based on four models: (a) structural invariance model (model 1), (b) weak factorial invariance model (model 2), (c) strong factorial invariance model (model 3) and (d) strict factorial invariance model (model 4). Population heterogeneity (PH), on the other hand, was tested based on two models, namely the variance equality model (model 5) and invariance of latent means model (model 6). As the reading comprehension, mathematical literacy and science literacy tests in PISA 2015 are each addressed as a single factor, the covariance invariance was not tested. Model comparisons in terms of both MI and PH were made within nested models. In other words, a step-by-step approach was assumed to evaluate the fit by comparing the more restricted nested model and the less restricted nested model.

With the correction of Satorra-Bentler, the χ^2 value (S-Bχ^2) was calculated owing to the presence of data deviating somewhat from the normal distribution and the calculation of the χ^2 difference degrees among the models. In numerous studies where analyses under the umbrella term SEM are performed, distribution(s) can deviate from the normal distribution within certain tolerance limits. In large sample sizes where a normal distribution cannot be obtained, S-Bχ^2 yields values close to χ^2 produced in normal distributions of small sample sizes. S-Bχ^2 is a considerably reliable test statistic in the evaluation of covariance structure models in various sample sizes and score distributions (Byrne, 2006; Everitt & Howell, 2005).

On the other hand, in hypothesis tests, the acceptable level of significance is .05; as n>300 in the data sets, except for the Turkish data, the cut off values for MI in multi-group CFA were as follows: ΔCFI<-.01 for the comparison of the three models, ΔSRMR≤.03 for the weak factorial invariance test, and ΔSRMR≤.01 for the strong factorial invariance and strict factorial invariance tests (Chen, 2007).

The Raju area index, which is a technique in the item response theory and compatible with the 2-parameter logistic model and Mantel-Haenszel technique, which is one of the classical differential item functioning approaches, was employed. The Mantel-Haenszel technique is a chi-square based test technique (Agresti, 2010). It measures the degree of difference in the performance of the focus and reference groups. In identifying the DIF, it can be stated that |Δ-MH | < 1 was an ignorable level (Level A); 1≤ |Δ-MH | < 1.5 a moderate level (Level B); |Δ-MH | ≥1.5 an important level in the item (Level C) (Zieky, 1993). In the present study, the values obtained from levels B and C were selected as DIF measures. In Raju's area indices method conducted with DIF, the item parameters or item characteristic curves obtained from the focus and reference groups are compared. If the item characteristic curves placed on the same plane overlap, or if the area between the item characteristic curves is zero, bias is believed to be nonexistent. As the size of the area between the item characteristic curves deviate from zero, the item bias increases (Raju, 1990). The significance level of the hypothesis test used in Raju's area index was accepted as .05.

3. Results

3.1. Findings Relation to Measurement Invariance and Population Heterogeneity

The MI for the groups taking the PISA reading comprehension test in English, Spanish, Arabic, Portuguese, Chinese, German, French and Turkish was tested via multi-group CFA. The findings obtained from the analyses are presented in Table 4.

Table 4

Measurement Invariance Findings Related to English, Spanish, Arabic, Portuguese, Chinese, German, French, and Turkish Groups of Reading Comprehension Test (Maximum Likelihood)

	S-Bχ^2(df)[1]	$\Delta\chi^2$(Δdf)	χ^2/df	$\Delta\chi^2$/Δd	CFI	ΔCFI	SRMR	ΔSRM	DECISIO

				f				R	N
English	1356.03(350)	–	3.87	–	.97	–	.031	–	–
Spanish	1581.78(350)	–	4.52	–	.95	–	.037	–	–
Arabic	685.90(350)	–	1.96	–	.94	–	.044	–	–
Portuguese	2348.02(350)	–	6.71	–	.86	–	.069	–	–
Chinese	675.18(350)	–	1.93	–	96	–	.041	–	–
German	866.32(350)	–	2.48	–	.94	–	.048	–	–
French	710.53(350)	–	2.03	–	.94	–	.052	–	–
Turkish	424.06(350)	–	1.21	–	.94	–	.070	–	–
Model 1[A]	15548.39(3192)	–	4.87	–	.88	–	.120	–	–
Model 2[B]	12559.25(2996)	2989.14(196)	4.19	15.25	.91	-.03	.084	.036	**H_0:Reject**
Model 3[C]	8694.29(2800)	3864.96(196)	3.11	19.72	.94	-.03	.069	.015	**H_0:Reject**
Model 4[D]	11442.94(2996)	-2748.65(-196)	3.82	14.03	.92	.02	.100	-.031	H_0:Accept
Model 5[E]	28843.04(3388)	-17400.10(-392)	8.51	44.39	.75	.17	.120	-.020	H_0:Accept
Model 6[F]	25803.95(3381)	3039.09(7)	7.63	434.16	.78	-.03	.160	-.040	**H_0:Reject**

[1] $p<.05$, [A] Configural Invariance (factor loads, factor correlation and error variance are constant), [B] Weak Factorial Invariance (factor loads, factor correlation and error variance are constant), [C] Strong Factor Invariance (factor loads and error variance are free; factor correlation is constant), [D] Strict Factorial Invariance (error variance is free; factor loads and factor correlation are constant), [E] Equality of variance, [F] Equality of latent means

When the fit indices are obtained as an outcome of the CFA analyses for each individual group taking the PISA reading comprehension test in the languages under study, it can be stated that the fit indices obtained from the eight groups meet the acceptable levels to a large extent. Accordingly, for the groups under study, the S-Bχ^2 and level of freedom values for all the groups, except English, Spanish and Portuguese, were found to be around 2 (it can be stated that as the number of individuals in the groups taking the English and Spanish tests is high, the S-Bχ^2/df for these groups turns out to be relatively high), while the CFI was above .90 and SRMR was below .08. It can be stated that among these data, the fit indices for Portuguese was not sufficient, but when the fit indices for the other languages were examined, each group under study was confirmed.

With the purpose of evaluating MI, initially the factor loadings, factor correlations and the structural model constructed with the hypothesis showed no significant difference in the error variances of the groups taking the reading comprehension test in English, Spanish, Arabic, Portuguese, Chinese, German, French, and Turkish. The analyses revealed that the S-Bχ^2 and degree of freedom ratios were below 5, CFI was below .90 and SRMR was above .08. The overall examination of the fit indices indicates that the fit indices of the structural model meet the acceptable level.

When the structural invariance (Model 1) and weak factorial invariance (Model 2) models were compared, it was observed that the fit improved with respect to the S-B$\Delta\chi^2$ and Δdf ratios of the two models. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 2321.13, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(196)=229.66$, $p<.05$. Moreover, it can be stated that there was a ΔCFI value of significant difference (>-.01), and that the ΔSRMR variance was also significant (≥.03).

When the weak factorial invariance (Model 2) and the strong factorial invariance (Model 3) models were compared, it was observed that the fit improved with respect to the S-B$\Delta\chi^2$ and Δdf ratios of the two models. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 3479.94, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(196)=229.66$, $p<.05$. Moreover, it can be stated that there was a ΔCFI value of significant difference (>-.01), and that the ΔSRMR variance was also significant (≥.01).

Finally, when the strong factorial invariance (Model 3) and the strict factorial invariance (Model 4) models were compared, it was observed that the fit weakened in terms of the S-B$\Delta\chi^2$ and Δdf ratios. The fit also weakened with respect to the ΔCFI and ΔSRMR values.

On the other hand, when Model 4 and Model 5 are compared with respect to the equality of variance, it can be stated that fit indices worsened to some degree in terms of the S-B$\Delta\chi^2$ and Δdf ratio. The fit also worsened with respect to the ΔCFI and ΔSRMR values. When Model 5 and Model 6 were compared in terms of the equality of latent means, it can be stated that fit is improved to some degree in terms of the S-Bχ^2 and df ratio and also improved with respect to the ΔCFI value. However, fit worsened in terms of the ΔSRMR value. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 834.19, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(7)=14.07$, $p<.05$.

The findings obtained from the multi-group CFA analyses in relation to the MI for the groups that took the PISA mathematical literacy test in English, Spanish, Arabic, Portuguese, Chinese, German, French and Turkish are presented in Table 5.

Table 5
Measurement Invariance Findings Related to English, Spanish, Arabic, Portoguese, Chinese, German, French, and Turkish Groups of Mathematical Literacy Test (Maximum Likelihood)

	S-Bχ^2(df)[1]	$\Delta\chi^2$(Δdf)	χ^2/df	$\Delta\chi^2$/Δdf	CFI	ΔCFI	SRMR	ΔSRMR	DECISION
English	1039.18(209)	–	4.97	–	.96	–	.033	–	–
Spanish	818.12(209)	–	3.91	–	.96	–	.034	–	–
Arabic	291.76(209)	–	1.40	–	.97	–	.045	–	–
Portuguese	409.36(209)	–	1.96	–	.96	–	.044	–	–
Chinese	462.97(209)	–	2.22	–	.96	–	.040	–	–
German	393.65(209)	–	1.88	–	.96	–	.040	–	–
French	404.01(209)	–	1.93	–	.95	–	.046	–	–
Turkish	223.01(209)	–	1.07	–	.99	–	.064	–	–
Model 1	19939.59(1980)	–	10.07	–	.70	–	.120	–	–
Model 2	13463.97(1826)	6475.62(154)	7.37	42.05	.81	-.11	.076	.044	**H_0:Reject**
Model 3	3930.96(1672)	9533.01(154)	2.25	61.90	.96	-.15	.044	.032	**H_0:Reject**
Model 4	6955.39(1870)	-3024.43(-198)	3.72	15.28	.92	.04	.086	-.042	H_0:Accept
Model 5	31276.07(2134)	-24320.68(-264)	14.66	92.12	.52	.40	.110	-.024	H_0:Accept
Model 6	29962.82(2127)	1313.25(7)	14.09	187.61	.54	-.02	.110	0	**H_0:Reject**

[1] $p<.05$

When the fit indices obtained from the CFAs performed separately for each group taking the mathematical literacy test in the languages under study are evaluated, it can be stated that the

fit indices for the eight groups meet the acceptable levels. Accordingly, it is observed that the S-Bχ^2 and degree of freedom ratios for the tests, except for those in Chinese, English and Spanish, were below 2 (for Chinese it was below 3; for English and Spanish it was below 5), CFI was equal to or above .95, and SRMR was below .08, except for Turkish. An overall examination of the fit indices indicates that the fit indices of the structural model meet the acceptable level. Moreover, the fit indices are confirmed separately for each group under study.

With the purpose of evaluating MI, initially the factor loadings, factor correlations and the structural model constructed with the hypothesis that there was no significant difference in the error variances of the groups taking the PISA mathematics test in English, Spanish, Arabic, Portuguese, Chinese, German, French, and Turkish were tested. The analyses revealed that the S-Bχ^2 and degree of freedom ratios were above 5, CFI was below .90 and SRMR was above .08. An overall examination of the fit indices indicates that the fit indices of the structural model do not meet the acceptable level.
When the structural invariance (Model 1) and weak factorial invariance (Model 2) models were compared, it was observed that fit improved in terms of the S-B$\Delta\chi^2$ and Δdf ratio for both models. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 10519.73, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(154)=183.86$, $p<.05$. Moreover, it can be stated that there was a significant difference in the ΔCFI value (>-.01); the difference in ΔSRMR was also found to be significant (≥.03).

When the weak factorial invariance (Model 2) and the strong factorial invariance (Model 3) models were compared, it was observed that the fit improved in terms of the S-B$\Delta\chi^2$ and Δdf ratio. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 12562.06, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(154)=183.86$, $p<.05$. Moreover, it can be stated that there was a significant difference in the ΔCFI value (>-.01); the difference in ΔSRMR was also found to be significant (≥.01).

Finally, when the strong factorial invariance (Model 3) and the strict factorial invariance (Model 4) models were compared, it was observed that the fit worsened in terms of the S-B$\Delta\chi^2$ and Δdf ratio. The fit also worsened with respect to the ΔCFI and ΔSRMR values.

On the other hand, when Model 4 and Model 5 were compared with respect to the equality in variance, it can be stated that fit worsened to some degree in terms of the S-B$\Delta\chi^2$ and Δdf ratio. The fit also worsened with respect to the ΔCFI and ΔSRMR values. When Model 5 and Model 6 were compared in terms of the equality in latent means, it can be stated that fit improved to some degree with respect to the S-B$\Delta\chi^2$ and Δdf ratio, and that fit did not change based on the ΔSRMR value. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 1578.03, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(7)= 14.07$, $p<.05$.

The findings obtained from the multi-group CFA analyses in relation to invariance among the groups that took the PISA science literacy test in English, Spanish, Arabic, Portuguese, Chinese, German, French and Turkish are presented in Table 6.

Table 6
Measurement Invariance Findings Related to English, Spanish, Arabic, Portuguese, Chinese,

German, French, and Turkish Groups of Science Literacy Test (Maximum Likelihood)

	S-Bχ^2(df)[1]	$\Delta\chi^2$(Δdf)	χ^2/df	$\Delta\chi^2$/Δdf	CFI	ΔCFI	SRMR	ΔSRMR	DECISION
English	2046.87(527)	–	3.88	–	.97	–	.033	–	–
Spanish	2489.98(527)	–	4.72	–	.95	–	.038	–	–
Arabic	890.27(527)	–	1.69	–	.95	–	.047	–	–
Portuguese	2567.35(527)	–	4.87	–	.89	–	.063	–	–
Chinese	1195.23(527)	–	2.27	–	.94	–	.045	–	–
German	991.28(527)	–	1.88	–	.95	–	.046	–	–
French	986.28(527)	–	1.87	–	.95	–	.051	–	–
Turkish	632.90(527)	–	1.20	–	.95	–	.076	–	–
Model 1	18444.89(4692)	–	3.93	–	.90	–	.099	–	–
Model 2	16439.10(4454)	2005.79(238)	3.69	8.43	.92	-.02	.077	.022	**H_0:Reject**
Model 3	11875.98(4216)	4563.12(238)	2.82	19.17	.95	-.03	.063	.014	**H_0:Reject**
Model 4	13686.42(4454)	-1810.44(-238)	3.07	7.61	.93	.02	.086	-.023	H_0:Accept
Model 5	27439.13(4930)	-13752.71(-469)	5.57	29.32	.84	.09	.093	-007	H_0:Accept
Model 6	24041.22(4923)	3397.91(7)	4.88	485.42	.86	-.02	.110	-.017	**H_0:Reject**

[1] $p<.05$

When the fit indices obtained from the CFAs performed separately for each group taking the science literacy test in the languages under study were evaluated, it can be stated that the fit indices for the eight groups meet the acceptable levels to a considerable degree. The analyses revealed that the S-Bχ^2 and degree of freedom ratios were below 5 for English, Spanish and Portuguese and below 3 for the other languages; CFI was found to be above .90 for Chinese, and equal to or above .95 for the other languages, except for Portuguese, for which it was below .90. As for SRMR, for Portuguese and Turkish, it was observed to be below .08, while for the other languages it was equal to or below .05. An overall examination of the fit indices indicates that the fit indices were confirmed separately for all the groups under study.

With the purpose of evaluating MI, initially the factor loadings, factor correlations and the structural model constructed with the hypothesis showed no significant difference in the error variances of the groups taking the PISA science literacy test in English, Spanish, Arabic, Portuguese, Chinese, German, French, and Turkish. The analyses revealed that the S-Bχ^2 and degree of freedom ratios were below 5, CFI was equal to .90 and SRMR was above .08. An overall examination of the fit indices indicates that the fit indices of the structural model meet the acceptable level at a moderate degree.

When the structural invariance (Model 1) and the weak factorial invariance (Model 2) models were compared, it was observed that the fit improved in terms of the S-B$\Delta\chi^2$ and Δdf ratio. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 2256.95, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(238)= 274.99$, $p<.05$. Moreover, it can be stated that there was a significant difference in the ΔCFI value (>-.01), while the difference in ΔSRMR was not found to be significant (<.03).

When the weak factorial invariance (Model 2) and the strong factorial invariance (Model 3) models were compared, it was observed that the fit improved in terms of the S-B$\Delta\chi^2$ and Δdf

ratio. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 90152.33, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(238)= 274.99$, $p<.05$. Moreover, it can be stated that there was a significant difference in the ΔCFI value (>-.01); the difference in ΔSRMR was also found to be significant (≥.01).

Finally, when the strong factorial invariance (Model 3) and the strict factorial invariance (Model 4) models were compared, it was observed that the fit worsened in terms of the S-B$\Delta\chi^2$ and Δdf ratio. The fit also worsened with respect to the ΔCFI and ΔSRMR values.

On the other hand, when Model 4 and Model 5 were compared with respect to equality in variance, it was observed that the fit worsened in terms of the S-B$\Delta\chi^2$ and Δdf ratio; it also worsened with respect to the ΔCFI and ΔSRMR values. When Model 4 and Model 5 were compared with respect to equality in latent means, it was found that fit improved to some degree with respect to the S-Bχ^2 and df ratio. Accordingly, it was found that the T_s value calculated for the S-Bχ^2 degree of difference was 1151.18, and that this value was higher than the critical value in the χ^2 distribution table, $\chi^2_{diff}(5)= 14.07$, $p<.05$. It can also be stated that fit improved with respect to the ΔCFI value, but worsened in terms of ΔSRMR.

The MI was analyzed once again within the scope of pairwise comparisons owing to the fact that MI could not be established for the commonly used languages in the PISA 2015 reading comprehension, mathematics, and science tests. Thus, the data for the original languages, English and French, were compared pairwise with Spanish, Arabic, Portuguese, Chinese and Turkish. The findings obtained as a result of the pairwise comparisons are presented in Table 7.

Table 7
Pairwise Comparisons of Measurement Invariance between English and French Forms and Other Languages

	Reading Comprehension		Mathematics		Science	
	English	French	English	French	English	French
Spanish	+	+	– (1,2,3)	– (1,2)	+	+
Arabic	– (1,2,3)	– (1,2,3)	– (1,2,3)	– (1,2,3)	– (1,2,3)	– (1,2,3)
Portuguese	– (1,2,3)	– (1,2)	– (1,2,3)	– (1,2,3)	– (1,2,3)	– (1,2)
Chinese	– (1,3)	+	– (1,2,3)	– (1,2)	+	+
German	+	+	+	+	+	+
Turkish	– (1,3)	– (1,2,3)	– (1,2,3)	– (1,2,3)	– (1,3)	– (1,2,3)

+ MI is provided; – MI is not provided; [1] there is a difference in χ^2_{diff}; [2] there is a difference in ΔCFI; [3] there is a difference in ΔSRMR.

As can be seen in Table 7, in all the tests MI was established only between German and the original languages. On the other hand, equivalence could not be obtained in any of the tests between the original languages and Arabic, Portuguese and Turkish. Moreover, equivalence was not obtained between Spanish and the original languages in the mathematical literacy test, nor was MI established between Chinese and the original languages in both the reading comprehension and mathematical literacy tests.

3.2. Findings Relation to DIF

The DIFs were evaluated according to the test forms belonging to the eight different languages based on the data sets of PISA 2015 science literacy, mathematical literacy and reading comprehension tests. In this evaluation, English and French were selected as reference languages as they are the languages with which the forms of the cognitive tests in the PISA implementation were developed. The related DIF findings are presented in Table 8.

Table 8
DIF Findings According to Form Language

	English						French					
	Mathematics		Science		Reading		Mathematics		Science		Reading	
	MH	Raju	MH	Raju	MH	Raju	MH	Raju	MH	Raju	MH	Raju
I1	1.36	1.19	.14	-1.75	**209.17**	**-9.93**	**7.65**	1.71	**7.82**	-1.51	**8.54**	**-2.45**
I2	4.27	**-2.27**	**6.57**	-1.82	**61.61**	**-6.05**	3.48	-1.35	**5.19**	-1.37	2.76	-1.13
I3	**30.17**	**-4.28**	**17.11**	**-3.12**	**4.35**	-1.89	2.44	-1.16	**9.27**	-1.79	.57	.72
I4	**8.47**	1.64	**85.91**	**-7.22**	**10.48**	**-2.75**	1.28	-1.34	**4.16**	-1.47	**12.33**	**-2.86**
I5	**7.62**	**3.13**	**5.60**	**3.27**	.94	**-2.17**	.26	**3.34**	.23	**2.10**	**28.29**	**-4.25**
I6	2.31	-1.65	3.69	**-3.45**	1.14	.70	**5.29**	**2.30**	**15.83**	**3.20**	**7.23**	**3.29**
I7	1.46	.71	**5.89**	-.98	**106.69**	**-8.00**	**6.04**	**1.97**	**23.25**	**5.30**	**4.80**	**-2.02**
I8	.52	**4.11**	.17	-.27	**18.70**	**4.57**	**50.83**	**5.32**	1.35	1.36	.15	-1.58
I9	**34.94**	**4.83**	.09	**2.55**	**21.15**	**-3.92**	.02	**3.24**	.40	-.94	**13.24**	**2.48**
I10	.10	**3.44**	**6.86**	1.21	**5.58**	**2.09**	.38	1.05	2.14	1.63	**6.70**	**-2.05**
I11	**97.71**	**2.95**	.57	-.33	**73.31**	**-7.50**	**10.58**	-1.16	**41.64**	**-5.19**	3.21	-1.68
I12	1.91	1.05	3.56	1.43	1.58	1.31	.40	.90	.00	-.34	**7.73**	**-2.21**
I13	3.19	.73	**9.21**	**2.56**	**37.78**	**-4.99**	.00	1.38	**5.61**	**2.28**	2.92	-1.69
I14	3.16	1.78	**26.31**	**4.13**	**18.60**	**-2.12**	.23	.87	**6.67**	**-1.99**	**4.59**	**2.08**
I15	.90	1.00	**10.68**	-1.73	**5.14**	-1.32	.66	-.30	1.44	1.18	**4.35**	**1.97**
I16	**4.17**	1.82	**18.47**	**-2.20**	**7.56**	**2.05**	1.21	**4.12**	.40	-1.63	3.34	-1.33
I17	2.13	1.87	.00	.62	**6.58**	**2.42**	.31	.19	1.52	**-2.10**	.05	1.13
I18	3.69	**2.88**	.24	-1.13	**6.06**	**2.60**	.01	1.41	3.19	-1.89	**4.37**	**2.22**
I19	**8.80**	**2.73**	**22.05**	**3.92**	.49	.74	.52	.75	2.44	-1.38	**8.08**	**-2.34**
I20	2.51	**4.31**	**44.98**	**6.81**	**21.82**	**4.56**	.04	**4.25**	**14.33**	**4.35**	**9.96**	**3.41**
I21	**24.68**	**-3.69**	**7.01**	**-2.95**	.27	-.43	.61	**3.61**	.48	-.87	**5.92**	**2.69**
I22	2.09	**-2.07**	1.79	.55	**210.59**	**-12.31**	.19	-.48	.70	.92	**5.30**	**2.55**
I23			.00	-1.85	.18	**-4.50**			**6.79**	**2.50**	3.73	**3.42**
I24			**5.52**	1.92	.73	**-2.39**			**3.90**	1.11	**7.40**	**-1.98**
I25			**5.25**	**2.35**	**82.87**	**7.39**			2.64	1.88	1.33	-.75
I26			**7.06**	-1.41	**23.40**	**2.44**			.42	-.86	**31.66**	**3.57**
I27			.70	.71	**67.75**	**6.57**			.48	-.93	**8.40**	**-1.96**
I28			**23.98**	**7.28**	.31	1.14			.12	.56	**5.35**	**-1.97**
I29			**8.79**	-1.42					2.99	-1.34		
I30			**42.16**	**-4.06**					**6.45**	-1.59		
I31			**21.15**	**-3.62**					.27	**2.58**		
I32			.02	**4.56**					1.96	1.73		
I33			**22.23**	**7.86**					**4.84**	**3.17**		
I34			2.22	**3.79**					**20.52**	1.73		

* Values shown in bold are DIF values based on .05 significance level. The others have no DIF.

When the DIF analysis findings were evaluated according to English, the reference test form of the translated test forms, DIF was detected in 14 (64%) of the 22 items on the mathematical literacy test, six of which are common (27%); in 26 items (76%) on the science literacy test, 15 (44%) of which are common, and in 23 items (82%) on the reading comprehension test, 18 (64%) of which are common. When the reference group findings of the French test form were evaluated, DIF was detected in 10 items (45%) on the mathematical literacy test, six of which are common (14%); in 18 items (53%) on the science literacy test, 8 (53%) of which are common, and in 20 items (71%) on the reading comprehension test, 19 (68%) of which are common. In the comparisons of both test forms, DIF was observed mostly in the reading comprehension items, while DIF was least observed in the items on the mathematical literacy test.

Whether or not translations from the fundamental languages of English and French into the languages of German, Spanish, Portuguese, Chinese, Turkish and Arabic led to DIF between these languages with reference to English and French was examined individually based on the data sets of PISA 2015 science literacy, mathematical literacy and reading comprehension tests. The related findings are presented in Table 9 and Table 10.

Table 9
DIF Findings According to English Form of Other Adaptation Language Forms

	Mathematical Literacy						Science Literacy						Reading Comprehension					
	G[1]	S[2]	P[3]	C[4]	T[5]	A[6]	G	S	P	C	T	A	G	S	P	C	T	A
I1	M[7]	-[9]	+[10]	M	-	+	-	+	R[8]	+	+	R	-	+	+	+	+	+
I2	-	R	R	M	R	R	-	R	R	M	R	R	M	+	+	+	+	+
I3	-	+	+	+	+	+	+	R	R	+	R	+	+	+	R	+	-	+
I4	+	R	R	+	R	R	+	+	+	+	R	+	+	+	+	+	R	R
I5	-	M	-	R	-	R	R	+	R	-	R	+	+	-	+	+	R	+
I6	+	R	-	+	-	R	R	+	R	-	R	R	+	+	R	M	-	-
I7	-	R	R	+	-	R	R	R	R	+	R	+	+	+	+	M	+	+
I8	-	-	+	-	-	+	R	R	R	M	R	R	+	+	+	+	R	+
I9	-	M	M	+	-	R	R	R	+	-	R	R	R	+	+	M	+	M
I10	M	+	-	+	-	R	R	+	R	-	R	R	-	+	+	+	-	-
I11	-	-	+	+	+	R	+	R	+	+	+	R	+	+	+	+	+	+
I12	+	R	R	+	-	R	-	-	+	+	R	R	+	-	+	+	+	-
I13	-	-	-	-	-	+	R	-	+	+	R	R	+	+	+	+	+	+
I14	+	-	-	+	-	R	R	+	+	+	R	R	+	+	M	+	-	+
I15	+	R	-	+	R	R	+	R	R	-	R	+	+	+	-	-	R	R
I16	+	-	-	-	M	+	R	+	+	R	R	R	-	R	+	+	+	+
I17	M	M	-	+	-	R	+	+	+	+	R	+	-	-	+	-	-	+
I18	-	-	-	+	-	-	+	+	-	+	R	R	+	-	-	-	R	-
I19	M	M	M	-	-	+	M	+	-	-	M	M	+	M	R	-	R	-
I20	-	-	R	+	-	R	+	+	+	-	+	+	-	+	+	+	R	+
I21	M	-	-	+	-	+	+	R	+	+	R	R	+	R	+	+	+	R
I22	R	R	M	+	-	R	R	R	-	+	+	R	+	+	+	+	+	+
I23							R	R	R	+	R	R	R	R	+	+	+	+
I24							-	-	-	+	+	-	+	R	+	+	+	+
I25							+	-	-	-	+	R	+	+	R	+	M	+
I26							+	+	R	-	R	R	-	R	+	+	+	+
I27							M	R	-	+	R	R	-	+	+	+	+	+
I28							-	+	+	R	R	R	+	M	+	+	-	M
I29							+	+	R	+	+	R						
I30							+	+	R	+	M	+						
I31							+	+	R	+	+	R						
I32							+	-	M	-	R	M						
I33							-	+	+	-	R	+						
I34							-	R	+	+	R	+						

[1] German, [2] Spanish, [3] Portuguese, [4] Chinese, [5] Turkish, [6] Arabic, [7] DIF was determined according to the Mantel-Haenszel, [8] DIF was determined according to the Raju's area index, [9] no DIF, [10] DIF was determined according to both methods

For the mathematical literacy items, when the DIF findings in which the English test form was taken as the reference group were evaluated, the number of items in which DIF was detected according to only one of the methods were six (27%) in German, eleven (50%) in Spanish, eight (36%) in Portuguese, three (14%) in Chinese, four (18%) in Turkish and fourteen (64%) in Arabic versions of the test. The total number of items with DIF determined according to at least one method were 12 (55%) in German, 13 (59%) in Spanish, 12 (55%) in Portuguese, 18 (82%) in Chinese, six (27%) in Turkish, and 21 (95%) in Arabic versions of the test. In mathematical literacy, the least number of items with DIF was in the Turkish language form, while the highest number of items with DIF was in the Arabic language form.

When the DIF analysis findings for the science literacy test were evaluated with respect to the English test form as a reference group, DIF was detected, based on just one method, in 13 items (38%) in German, in 12 items (35%) in Spanish, in 15 items (44%) in Portuguese, in four items (12%) in Chinese, in 25 items in Turkish (74%) and in 23 items in Arabic (68%) versions of the test. The total number of items detected with DIF according to at least one method is 27 (79%) in German, 29 (85%) in Spanish, 28 (82%) in Portuguese, 23 (68%) in Chinese, 34 (100%) in Turkish and 33 (97%) in Arabic versions of the test. The least number of items with DIF in science literacy was in the Chinese language form, while the highest number of items with DIF was in the Turkish language form. When the DIF findings in which the English test form is taken as the reference group for the reading comprehension items were evaluated, the items detected with DIF according to only one of the methods were three (11%) in German, seven (25%) in Spanish, five (18%) in Portuguese, three (11%) in Chinese, eight (29%) in Turkish and five (18%) in Arabic versions of the test. The total number of items with DIF detected according to at least one method was 21 (75%) in German, 24 (86%) in Spanish, 26 (93%) in Portuguese, 24 (86%) in Chinese, 22 (79%) in Turkish and 23 (82%) in Arabic versions of the test. The least number of items with DIF for the reading comprehension test was the German language form, while the highest number of items with DIF was the Portuguese language form.

Table 10
DIF Findings According to French Form of Other Adaptation Language Forms

	Mathematical Literacy						Science Literacy						Reading Comprehension					
	G	S	P	C	T	A	G	S	P	C	T	A	G	S	P	C	T	A
I1	-	-	-	+	M	+	-	-	R	+	+	R	-	+	+	+	+	+
I2	-	R	R	-	-	R	M	R	R	-	R	R	-	-	+	R	-	+
I3	-	+	R	R	+	+	R	+	+	-	R	R	-	-	-	M	-	-
I4	-	R	R	-	R	R	+	+	+	+	R	+	-	+	+	+	+	+
I5	-	+	-	R	-	R	-	-	-	R	R	R	R	+	+	+	+	R
I6	+	-	-	+	-	R	+	+	R	+	+	+	+	R	+	-	M	-
I7	-	R	+	+	-	R	+	M	M	+	+	+	R	-	R	+	R	R
I8	+	+	+	+	-	+	-	-	-	+	R	R	R	+	+	-	R	+
I9	-	-	-	+	-	R	R	R	+	-	R	R	-	M	M	+	M	M
I10	-	+	-	+	-	R	R	+	R	-	R	R	+	+	+	R	-	M
I11	-	-	R	+	+	R	+	+	+	+	+	+	-		+	+	+	-
I12	-	-	-	R	-	R	-	-	R	-	R	R	-	+	+	R	R	-
I13	-	-	-	-	-	R	R	M	+	-	+	+	+	R	+	R	R	-
I14	-	-	-	R	-	R	R	+	+	+	R	R	-	-	-	+	M	-
I15	-	-	-	+	R	R	+	-	-	M	R	R	-	-	M	M	+	R
I16	M	R	R	R	+	+	R	R	+	-	R	R	-	-	+	R	+	-
I17	-	M	-	+	-	R	R	+	+	+	R	R	-	-	-	-	-	+
I18	-	-	-	M	-	-	+	R	-	+	+	+	+	R	-	M	-	-
I19	-	-	-	+	-	R	-	+	-	-	M	M	+	+	+	M	+	M
I20	-	R	R	+	-	R	-	+	-	+	R	+	-	+	+	+	-	+
I21	M	R	R	+	M	M	+	R	R	-	R	R	+	-	+	+	R	R
I22	-	-	-	+	-	R	-	-	-	+	R	R	+	+	+	+	+	+
I23							+	-	M	+	+	R	M	+	+	-	+	R
I24							M	-	-	-	R	M	M	+	+	+	-	R
I25							-	-	-	-	R	R	-	-	+	+	-	R
I26							M	R	-	-	R	R	+	+	+	+	+	+
I27							+	-	-	+	R	R	-	+	+	+	+	+
I28							-	-	-	-	R	R	M	-	+	+	R	M
I29							R	-	M	+	+	+						
I30							-	-	+	+	M	R						
I31							-	-	-	-	+	R						
I32							R	-	+	-	R	M						
I33							-	+	+	-	R	+						

I34	M	+	+	-	R	R

For the mathematical literacy items, when the DIF findings in which the French test form were taken as the focus group was evaluated, the number of items in which DIF was detected according to only one of the methods were two (9%) in German, seven (32%) in Spanish, seven (32%) in Portuguese, six (27%) in Chinese, four (18%) in Turkish and 17 (77%) in Arabic versions of the test. The total number of items detected with DIF according to at least one method is four (18%) in German, 11 (50%) in Spanish, nine (41%) in Portuguese, 19 (86%) in Chinese, seven (32%) in Turkish and 21 (95%) in Arabic versions of the test. The least number of items with DIF in mathematical literacy was found to be in the German language form, while the highest number of items with DIF was in the Arabic language form. When the DIF findings in which the French test form is taken as the reference group for the science literacy items, the number of items with DIF detected based on only one of the methods was 13 (38%) in German, eight (24%) in Spanish, nine (26%) in Portuguese, two (6%) in Chinese, 25 (74%) in Turkish and 25 (74%) in Arabic versions of the test. The total number of items with DIF detected according to at least one method was 22 (65%) in German, 19 (56%) in Spanish, 21 (62%) in Portuguese, 17 (50%) in Chinese, 34 (100%) in Turkish and 34 (100%) in Arabic versions of the test. The least number of items with DIF in science literacy was found to be in the Chinese language form, while the highest number was in the Turkish and Arabic language forms. When the DIF findings in which the French test form is taken as the reference group for the reading comprehension items, the number of items with DIF detected based on only one of the methods was six (21%) in German, four (14%) in Spanish, three (11%) in Portuguese, nine (32%) in Chinese, nine (32%) in Turkish and 11 (39%) in Arabic versions of the test. The total number of items with DIF detected according to at least one method was 14 (50%) in German, 17 (61%) in Spanish, 24 (86%) in Portuguese, 24 (86%) in Chinese, 20 (71%) in Turkish and 20 (81%) in Arabic versions of the test. The least number of items with DIF in reading comprehension was found to be in the German language form, while the highest number was in the Portuguese and Chinese language forms.

In the condition where the French was taken as the reference group, it was found that there were fewer items with DIF when compared to the DIF findings in which English was taken as the reference group. According to the French test language, the total number of items with DIF was 337: 71 in the mathematical literacy test, 147 in the science literacy test and 119 in the reading comprehension test. According to the English test language, the total number of items with DIF was 396: 82 in the mathematical literacy test, 174 in the science literacy test and 140 in the reading comprehension test.

4. Conclusion and Discussion

In test development and adaptation studies in the field of education, it is important to psychometrically test the invariance of the factor design obtained from the exploratory and/or confirmatory factor analyses, which are the most frequently used methods, for the defined groups in order to obtain empirical evidence in terms of structural validity. Researchers frequently make between-group comparisons in order to produce either theoretical or practical knowledge, and they naturally want the decisions they make about the population they want to generalize their study findings on to be as free from error as possible. As these comparisons are generally made based on the scores obtained from tests, the factor structure of the tool to be used in the comparison must definitely be examined by the researchers in order to determine whether or not it is equal for all the groups. The reason is that if the defined factor structure of the measurement tool is not equal for all the groups, the scores that the groups

obtain from these structures will not express the same meaning. When the factor structures are made equal for all of the groups, it can be maintained that the factor design under study will have the same meaning for all the groups; thus, the scores the groups obtain from the tests will be valid.

The general aim of the PISA implementation, which is accepted as the largest education study in the world, is not to focus on students' individual achievements, nor is it to provide them with feedback. Conversely, it provides the participant countries with a report of an internal and external comparative evaluation of the success of their education system. To this end, in PISA, three fundamental skills are tested: reading comprehension skill, mathematical literacy, and science literacy. The tests on which the assessments are based are developed in English and French. After the tests are developed, they are adapted to other languages. At this point, the factor structures of the tests developed in the original language and the test forms adapted to the target language should be equal (MI). Thus, in the present study, whether or not MI is established in the tests measuring the three fundamental skills in terms of the commonly used languages in subject, and even if it was not established, the variance item functioning based on the language variable was examined.

The best working model among the four models for reading comprehension skills with respect to the MI analyses was the strong factorial invariance model. Accordingly, the forms of the PISA reading comprehension test in different languages was considered to possess invariance in terms of factor designs. In other words, it was considered that MI was not established. When the obtained findings were examined in terms of PH, it was deduced that the fit indices of the equality model varied from the fit indices of the variance equality model. In each case, due to the weak fit of the models, the PH could not be obtained. Even if the covariance matrices of the reading comprehension test are considered to be equal in terms of the eight languages, the fact that the fit indices in the configural invariance model do not generally meet the accepted levels increases doubts regarding the lack of MI. It was determined via the pairwise comparisons that invariance was established between the original languages themselves and additionally between the original languages and the languages of Spanish and German. With respect to Chinese, invariance is established only with French, but not with English. At this point, it is interesting that invariance is not established in ay condition between the original languages and the languages of Arabic, Portuguese and Turkish.

The best working model for mathematical literacy in terms of MI is again the strong factorial invariance model. Accordingly, it has been accepted that the factor designs of the forms in different languages of the mathematical literacy test are not equal; that is, the MI is not established. When the obtained findings are evaluated in terms of PH, it can be deduced that the fit indices in the equality model of latent means varied from the fit indices of the variance equality model. However, in every condition, the model fits were considerably weak, and in this condition, PH could not be established once again. The fact that the covariance matrices of the mathematical literacy test were not equal in terms of the eight languages, and that the fit indices in the configural invariance model did not meet the acceptable levels again increases doubts regarding lack of MI. The outcome of the pairwise comparisons conducted with respect to mathematical literacy is quite striking. Accordingly, it was found that invariance was established within the original languages themselves and in addition only between the original languages and German, but that invariance was not established in any condition between the original languages and all the other languages.

Finally, the best working model among the four models for science literacy with respect to the

MI analyses was the strong factorial invariance model. Accordingly, it was accepted that the factor designs of the different language forms of the science literacy test were not equal in terms of factor designs. When the obtained findings were evaluated in terms of PH, it was stated that the results obtained for reading comprehension and mathematics were consistent; in other words, PH could not be established. In the pairwise comparisons, it was found that invariance was established among the original languages themselves and additionally between the original languages and Spanish, Chinese and German. Similar to the results obtained from reading comprehension and mathematics, it can be stated that there was no equality between the factor structures of the forms in the original languages and those in Arabic, Portuguese and Turkish.

When the literature on the research topic within the scope of large-scale tests is examined, only a limited number of research studies are encountered. In a study by Ercikan and Koh (2005), the MI was tested for the mathematics and science tests of the TIMSS 1995 implementation. It was concluded in the study that MI could not be established in any of the eight booklets in the comparison of the USA-English and French languages in the mathematics test; similarly, MI could not be established in three of the eight booklets in the comparison of French and Canadian and British English. Furthermore, in the science test, it was found that MI could not be established in any of the eight booklets in the comparison between USA-English and French, and between French and Canadian and British English. In another study on the 1999 TIMSS mathematics scores, Wu, Li and Zumbo (2007) looked into the MI among the scores in 21 countries. It was found that MI in two of the 21 comparisons were established; however, it was concluded that among 19 of the comparisons, weak or strong factorial invariance models possessed better fit indices. Ayvallı (2016) reported that the best working model for MI for the PISA 2012 mathematical literacy test among the OECD countries was the strong factorial invariance model.

In the present study in which DIF was examined, it was determined that in the condition where French was taken as the reference group, there were fewer items with DIF when compared to conditions where English was taken as the reference group. This shows similarity with the findings reported by Alatlı and Bökeoğlu (2018). In their study where the French test form was taken as the reference group, Alatlı and Bökeoğlu (2018) identified items with DIF in mathematical literacy with percentages of 40% for Chinese, 32% for Turkish; in science literacy, 0% for Chinese and 20% for Turkish and in reading comprehension, 21% for Chinese and 7% for Turkish. In similar comparisons where English was taken as the reference group, items with DIF were found at ratios ranging between 21% and 67%.

In the present study where two languages were taken as reference groups, and a total of eight languages were compared, DIF was detected in more than half of the items. This situation can be interpreted as the variety in specifically the number of languages leading to an increase in items with variance item functioning.

In comparisons to be made in large-scale international test implementations such as PISA and TIMSS, whether or not items display DIF should be examined, and when items do display DIF, the underlying reasons should be investigated. In addition, to eliminate problems deriving from translation and cultural factors, the people in international commissions should be experts and should work carefully. Care should be given to the selection of translators who work in such kinds of exams and to the cooperative work of translators and experts on assessment and evaluation.

In conclusion, that MI could not be obtained in all the tests in terms of the commonly used languages and because the number of items with DIF turned out to be high, it can be asserted that language and cultural invariance could not be ensured. Thus, it is essential to approach the decisions made based on the comparisons among countries and the rankings based on the test implementations in different languages, which constitutes the foundation of the implementation, with doubt. The translations from the original language to the target language and adaptation processes should be questioned as a potential source of this condition. In these studies, not only language experts, but also people who know the language and culture very well, experts on educational psychology, experts on the field of assessment and evaluation and experts on testing should be made to take an active role in the process.

There are two limitations in this study. Firstly, only eight languages were examined: seven are widely used languages, and the eighth is Turkish. Secondly, only two DIF detection methods are used. Consequently, the results are confined to those languages. In future studies, researchers can repeat the research using other languages and different DIF detection methods to identify issues in standardized testing arising from culture and translation and to outline ways to improve test accuracy. Because standardized testing plays such a large role in creating curricula, understanding the areas where improvements are necessary is essential to decision making in education; therefore, it would be beneficial for future studies to address the limitations and broaden the scope of the present study.

References

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4-14. https://doi.org/10.3102/0013189X033001004

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1-28. https:// doi.org/10.3102/00346543074001001

Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons. https://doi.org/10.1002/ 9780470594001

Alatlı, B. K., & Bökeoğlu, Ö. Ç. (2018). Uluslararası öğrenci değerlendirme programı (PISA -2012) okuryazarlık testlerinin ölçme değişmezliğinin incelenmesi. *İlköğretim Online*, *17*(2), 1096-1115. DOI: 10.17051/ ilkonline.2018.419357

Allalouf, A. (2000, April). *Retaining Translated Verbal Reasoning Items by Revising DIF Items*. Annual Meeting of the American Educational Research Association, New Orleans, LA, USA. https://doi.org/ 10.1111/j.1745-3984.1999.tb00553.x

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of educational measurement*, *36*(3), 185-198.

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

Ayvallı, M. (2016). *PISA 2012 matematik okuryazarlığı testinin ölçme değişmezliğinin incelenmesi* (Yayınlanmamış Yüksek Lisans Tezi). Akdeniz Üniversitesi, Antalya.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. (First Edition). New York: Guilford Publications, Inc.

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004, December). *Academic English in Fifth-Grade Mathematics, Science, and Social Studies Textbooks*. Center for Research on Evaluation Standards and Student Testing CRESST. (ERIC Document Reproduction Service No. ED483409). https://files.eric.ed.gov/fulltext/ED483409.pdf

Byrne, B. M. (2006). *Structural equation modeling with EQS and EQS/Windows: Basic*

concepts, applications, and programming. (Second Edition). California: Sage Publications, Inc.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14* (3), 464-504. https://doi.org/10.1080/10705510701301834

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23-35. https://doi.org/10.1207/s15327574ijt0501_3

Everitt, B. S. & Howell, D. C. (2005). *Encyclopedia of statistics in behavioral sciences*. Chichester: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192

French, B. F., Finch, W. H., & Vazquez, J. A. V. (2016). Differential item functioning on mathematics items using multilevel SIBTEST. *Psychological Test and Assessment Modeling*, *58*(3), 471.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in educational evaluation*, *33*(1), 69-86. https://doi.org/10.1016/j.stueduc.2007 .01.006

He, W., & Wolfe, E. W. (2010). Item equivalence in English and Chinese translation of a cognitive development test for preschoolers. *International Journal of Testing*, *10*(1), 80-94. https://doi.org/10.1080/ 15305050903534738

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*(2), 378-390. https://doi.org/10.1080/01443410.2014.946890

Huang, X. (2010). *Differential Item Functioning: The Consequence of Language, Curriculum, or Culture?*. University of California, Berkeley.

Marotta, L., Tramonte, L., & Willms, J. D. (2015). Equivalence of testing instruments in Canada: Studying item bias in a cross-cultural assessment for preschoolers. *Canadian Journal of Education*, *38*(3), 1.

Nguyen, H. T., & Cortes, M. (2013). Focus on Middle School: Teaching Mathematics to ELLs: Practical Research-Based Methods and Strategies: Detra Price-Dennis, Editor. *Childhood Education*, *89*(6), 392-395. https://doi.org/10.1080/00094056.2013. 854130

Nunnaly, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. (Third Edition). New York: McGraw-Hill, Inc.

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, *13*(3), 272-293. https://doi.org/10.1080/ 15305058.2012.738266

Price, L. R., & Oshima, T. C. (1998). Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207. https://doi.org/10.1177/ 014662169001400208

Roth, W. M., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *The Journal of Mathematical Behavior*, *40*, 88-105. https://doi.org/10.1016/j.jmathb.2015.01.004

Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European journal of psychological assessment*, *13*(1), 29 37. https://doi.org/10.1027/1015-5759.13.1.29

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach

and also an individual differences approach: an international testing context. *Psychology Science*, *50*(3), 403.

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