AN INVESTIGATION INTO THE MEASUREMENT INVARIANCE OF
PISA 2012 MATHEMATICAL LITERACY TEST

Merve Ayvalli,
Bayram Biçak
Akdeniz University, Educational Faculty,
Department of Educational Sciences, Turkey

Abstract:
The aim of this study is to investigate the measurement invariance of PISA 2012 mathematical literacy among the OECD member countries, and gender and region groups in Turkey. Among cognitive test booklets implemented in PISA 2012, booklet 8 which was used commonly by all countries was selected for this correlational survey study. The study was conducted using the dataset belonging to 23,311 students that took booklet 8 in the OECD member countries and 377 students that took this booklet in Turkey. Initially, measurement models were verified for all groups by performing a confirmatory factor analysis separately for OECD member countries, gender and region groups. The research then proceeded with the phase of testing the measurement invariance by testing the equivalence of covariance matrices for all groups. The measurement invariance was tested by means of a multi-group confirmatory factor analysis. The results revealed that the measurement invariance held true for the gender and region groups in Turkey, but not for the OECD member countries, and that the strong factorial invariance model was the one that worked most successfully among the models of measurement invariance.

Keywords: PISA 2012, mathematical literacy, measurement invariance, multi-group confirmatory factor analysis

1. Introduction

In the globalized world, international initiatives have also gained importance in addition to national assessment and evaluation initiatives which are effective in identifying and reformulating education policies. Programme for International Student Assessment (PISA), carried out by the Organization for Economic Cooperation and Development (OECD) is one of the exams which is internationally implemented and
regarded as a reference by many countries to identify at the international level the position of countries and determine their differences and shortcomings compared to other countries in terms of educational systems. PISA implementations conducted every three years to test three basic domains - mathematical literacy, science literacy and reading skills- feature one of these three main domains at each implementation period. Mathematical literacy was the focus of PISA 2012. OECD (2013) defined mathematical literacy as a domain that assists individuals to become aware of their roles in the world and assists constructive, committed and reflective citizens in making sound judgments and decisions. This definition states that mathematical literacy includes the processes of formulating situations mathematically, applying mathematical concepts, methods, factors and tools to mathematical thinking and interpreting mathematical outputs. The mathematical performance scores of countries changed between 368 and 613 in the implementation (OECD, 2013) to which 65 countries participated. While the mean score of OECD member countries was found to be 494, mean score for all countries was 487. Turkey’s mean score for mathematical literacy in PISA 2012 was 448 and this score was included at the second level (OECD, 2014). Two main points that affect students’ mathematical literacy performances in the Turkish sample are noteworthy. The first of these points is the difference between female and male students’ performances and the condition termed as gender gap. Gender gap was identified in a large majority of the countries- including Turkey- that are OECD members and non-OECD countries. The second point is related to the region-based differences of performance in Turkey. Score differences were observed between regions with the lowest and highest means (Ministry of National Education, 2015).

While PISA results are highly effective for many participating countries to assess their educational systems and identify educational policies, PISA allows comparison of many groups and countries. Measurement between groups includes psycho-metric assessment procedures regarding personality, intelligent or other constructs and the measurement tool that will assess a specific psychological construct should be made meaningful for the target group or cultures in order to have equal and comparable results for the groups with gender differences, regional or cultural differences. To apply the measurement results conducted by tools with proven reliability and validity on different groups and to compare these results will only be possible through interpretation of the measured psychological characteristic in the same way by all groups. Today, the fact that there is an increasing need for different measurement tools to assess different characteristics and the fact that there is an abundance of studies on comparative studies based on different cultures, regions, cities and various demographic characteristics have led to the requirement that the results obtained from using the measurement tool in a new universe must be interpreted by taking the effects generated by differences between groups into consideration (Mushquash and Bova, 2007). This is crucial to ensure the generalization of the measured construct to different groups (Brown, 2006). Complying with the statement that “no differences are generated by the measurement tool” in measurements that utilize the same measurement tool in different groups may cause shortcomings in comparisons between groups and relevant
interpretations. If this compliance can be validated, comparisons and analyses become meaningful. Otherwise, obtained results will lose their meaning since the reliability and validity of the results will be compromised (Baştuta, 2010). Measurement invariance studies are conducted to determine to what extent the measurement models can be generalized between groups and against time. If a measurement tool has been developed to for implementation on a heterogeneous group, it is necessary to prove that the measurement characteristics of the test are equal at sub groups. Measurement equivalence is mostly related to the measurement tool itself rather than to the characteristics of the individuals included in the universe to which measurement procedures are applied. The purpose of measurement invariance studies is not developing a new measurement tool. Measurement invariance is used to allow comparisons between different groups (Cheung and Rensvold, 2000).

Measurement invariance is defined as the formal assessment of the equivalence of psycho-metric qualities of a psychological measurement tool such as reliability and construct validity in different groups (Herdman, Fox-Rushby and Badia, 1998). Byrne and Watkins (2003) define measurement invariance as perceiving and interpreting the items in a measurement tool completely in the same manner. Ensuring the validity of measurements for groups who are compared is the basis of measurement invariance (Tyson, 2004).

Examination of literature shows that measurement invariance studies are specifically conducted to explore the generalizability psychological constructs between cultures and to determine the comparability of these constructs. In addition, measurement invariance is also tested in groups with the same culture for comparisons conducted for sub groups such as gender, age and ethnic origin.

In this context, the problem statement of the study is related to testing measurement invariance for mathematical literacy in PISA 2012 for gender, region and OECD-member countries. This study sets out to discuss whether the results are comparable between groups by testing the measurement invariance of mathematical literacy in PISA 2012 for gender, region and OECD-member countries. With this aim in mind, answers were sought to the questions provided below:

1) Does measurement invariance of mathematical literacy based on gender hold true for the data set obtained as a result of PISA 2012 Turkey implementation?
2) Does measurement invariance of mathematical literacy based on regions hold true for the data set obtained as a result of PISA 2012 Turkey implementation?
3) Does measurement invariance of mathematical literacy based on OECD member countries hold true for the data set obtained as a result of PISA 2012 Turkey implementation?

2. Material and Methods

2.1. Research Model
This study employed relational screening model since it aimed to determine the validity level of the cognitive test in PISA 2012 assessment.
2.2. Universe and Sample

PISA 2012 included approximately 510,000 15-year old students from a total of 65 countries—34 OECD member countries and 31 non-OECD countries. In Turkey, a total of 4848 students from 170 schools random sampling selected by PISA international center via random sampling participated in PISA to represent the 955,349 15-year olds. The universe of the study was composed of 15 year old students that participated in PISA 2012 form OECD member countries.

In each country, the cognitive test items used in PISA 2012 were implemented by distributing them to 13 separate booklets (MoNE, 2015). Questions included in each booklet were different from one another therefore booklet 8, which was mutually used in all countries, was selected for the study and analyses were conducted on the items that aimed to test mathematical literacy.

This study identified two separate universes based on the purpose of the research since it included all the individuals that were given booklet 8 in OECD member countries. The first universe defined according to sub goals was composed of 23,311 students that were given booklet 8 in OECD member countries and the second universe was composed of the 377 students that were given booklet 8 in Turkey.

This study also aimed to conduct the invariance study in Turkey on gender and regions, however, when the number of participants distributed to 12 regions based on Nomenclature of Territorial Units for Statistics did not meet the assumptions for the analysis; the regions were combined via expert views obtained from human and physical geography fields and the number of regions was reduced to three. Socio-economic situations of the regions, their educational statuses and cultural characteristics were taken into account when combining the regions for the study. In this context, Aegean, Mediterranean and West Anatolian regions were combined to form “Region 1; Central Anatolia, Southeastern Anatolia, Middle Eastern Anatolia, North-Eastern Anatolia and Eastern Black Sea regions were combined to form “Region 2” and Istanbul Western Marmara, Eastern Marmara and Western Black Sea regions were combined to form “Region 3”. Table 1 presents the gender based distribution of individuals that received booklet 8 in Turkey:

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>183</td>
<td>48.5</td>
</tr>
<tr>
<td>Male</td>
<td>194</td>
<td>51.5</td>
</tr>
<tr>
<td>Total</td>
<td>377</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2 displays the region based distribution of individuals that received booklet 8 in Turkey.
Table 2: Distribution of the sample according to regions

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>141</td>
<td>37.4</td>
</tr>
<tr>
<td>Region 2</td>
<td>109</td>
<td>29</td>
</tr>
<tr>
<td>Region 3</td>
<td>127</td>
<td>33.6</td>
</tr>
<tr>
<td>Total</td>
<td>377</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3 shows the country based distribution of individuals that received booklet 8 in OECD member countries.

Table 3: Distribution of the universe according to OECD member countries

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1198</td>
<td>5.13</td>
</tr>
<tr>
<td>Austria</td>
<td>373</td>
<td>1.60</td>
</tr>
<tr>
<td>Belgium</td>
<td>661</td>
<td>2.83</td>
</tr>
<tr>
<td>Canada</td>
<td>1679</td>
<td>7.20</td>
</tr>
<tr>
<td>Chile</td>
<td>532</td>
<td>2.28</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>429</td>
<td>1.84</td>
</tr>
<tr>
<td>Denmark</td>
<td>558</td>
<td>2.39</td>
</tr>
<tr>
<td>Estonia</td>
<td>404</td>
<td>1.73</td>
</tr>
<tr>
<td>Finland</td>
<td>665</td>
<td>2.85</td>
</tr>
<tr>
<td>France</td>
<td>353</td>
<td>1.51</td>
</tr>
<tr>
<td>Germany</td>
<td>387</td>
<td>1.66</td>
</tr>
<tr>
<td>Greece</td>
<td>391</td>
<td>1.67</td>
</tr>
<tr>
<td>Hungary</td>
<td>361</td>
<td>1.54</td>
</tr>
<tr>
<td>Iceland</td>
<td>258</td>
<td>1.10</td>
</tr>
<tr>
<td>Ireland</td>
<td>376</td>
<td>1.61</td>
</tr>
<tr>
<td>Israel</td>
<td>561</td>
<td>2.40</td>
</tr>
<tr>
<td>Italy</td>
<td>2643</td>
<td>11.33</td>
</tr>
<tr>
<td>Japan</td>
<td>477</td>
<td>2.04</td>
</tr>
<tr>
<td>Korea</td>
<td>394</td>
<td>1.69</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>403</td>
<td>1.72</td>
</tr>
<tr>
<td>Mexico</td>
<td>2591</td>
<td>11.11</td>
</tr>
<tr>
<td>New Zealand</td>
<td>346</td>
<td>1.48</td>
</tr>
<tr>
<td>Norway</td>
<td>351</td>
<td>1.50</td>
</tr>
<tr>
<td>Poland</td>
<td>380</td>
<td>1.63</td>
</tr>
<tr>
<td>Portugal</td>
<td>442</td>
<td>1.89</td>
</tr>
<tr>
<td>Slovakia</td>
<td>396</td>
<td>1.69</td>
</tr>
<tr>
<td>Slovenia</td>
<td>453</td>
<td>1.94</td>
</tr>
<tr>
<td>Spain</td>
<td>1963</td>
<td>8.42</td>
</tr>
<tr>
<td>Sweden</td>
<td>354</td>
<td>1.51</td>
</tr>
<tr>
<td>Switzerland</td>
<td>864</td>
<td>3.70</td>
</tr>
<tr>
<td>Turkey</td>
<td>377</td>
<td>1.61</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>954</td>
<td>4.09</td>
</tr>
<tr>
<td>United States of America</td>
<td>414</td>
<td>1.77</td>
</tr>
</tbody>
</table>

2.3. Research Data
Booklet 8 which was implemented in all countries was selected in this study and since scoring was differentiated according to item types, only 11 multiple choice items that
were related to mathematical literacy were included in the analysis. These items were scored “1” point for correct answers and “0” for incorrect answers. Table 4 provides the items used in the study. The data were accessed via OECD’s official website in the PISA section and downloaded from the following link: https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm.

Table 4: Mathematical literacy items included in the analysis

<table>
<thead>
<tr>
<th>Items</th>
<th>Mathematical literacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM00KQ02-Wheelchair Basketball Q2</td>
<td></td>
</tr>
<tr>
<td>PM906Q01- Crazy Ants Q1</td>
<td></td>
</tr>
<tr>
<td>PM915Q01- Carbon Tax Q1</td>
<td></td>
</tr>
<tr>
<td>PM915Q02- Carbon Tax Q2</td>
<td></td>
</tr>
<tr>
<td>PM982Q01- Employment Data Q1</td>
<td></td>
</tr>
<tr>
<td>PM982Q02- Employment Data Q2</td>
<td></td>
</tr>
<tr>
<td>PM982Q03T- Employment Data Q3</td>
<td></td>
</tr>
<tr>
<td>PM982Q04- Employment Data Q4</td>
<td></td>
</tr>
<tr>
<td>PM992Q01- Spacers Q1</td>
<td></td>
</tr>
<tr>
<td>PM992Q02- Spacers Q2</td>
<td></td>
</tr>
<tr>
<td>PM992Q03- Spacers Q3</td>
<td></td>
</tr>
</tbody>
</table>

2.4. Data Analysis

Before data analysis, all premises were examined and organized according to the analyses that would be conducted. The psychometric characteristics of the data set were explored and coefficients of kurtosis and skewness and KR-20 reliability coefficients were calculated. Confirmatory factor analysis was conducted to obtain proof of construct validity. Then, equivalence of covariance matrices was tested and measurement invariance of the items was examined according to OECD member countries, regions in Turkey and gender by using multi-group confirmatory factor analysis. The analyses were undertaken with the help of LISREL 8.7 program and maximum likelihood estimation method was utilized to estimate the model parameters over asymptotic covariance matrix.

In order to undertake analyses, data set were tested to observe whether basic assumptions were met. Results related to these assumptions are provided below.

Missing Values: Since it is believed in the study that missing values would not significantly affect sample size, they were excluded from the data set.

Extreme Values: One way extreme value screening was done in the study and cut point/breakpoint was determined as ±3 (Raykov and Marcoulides, 2008) and the values outside this point were excluded from the analysis. The analysis was conducted for 23,311 individuals in OECD member countries and 377 individuals in Turkey.

Normality: since it is difficult to calculate multi variable normality in the framework of Structural Equation Modeling, the normality of the data was determined by examining kurtosis and skewness coefficients, methods of calculating single variable normality (Weston and Gore, 2006). Obtained results are presented in the findings section.
Multi-collinearity Problem: Correlations between independent variables were examined for multi-collinearity problem and it was found that correlations for variables changed between .046 and .31 tolerance values changed between .78 and .94. It can be argued that multi-collinearity problem does not exist in cases where tolerance values are bigger than .01 (Tabachnick and Fidell, 2013). When the obtained values are considered, it is possible to state that multi-collinearity problem does not exist.

In this framework, it can be argued that all premises are met.

While evaluating the results of multi-group confirmatory factor analysis results, S-B $\chi^2$ was used as $\chi^2$ value since S-B $\chi^2$ correction decreases the effect of sample size on $\chi^2$ in data sets with different sample sizes and score distribution and $\chi^2$, $\chi^2$/sd, CFI and SRMR fit indices were taken into consideration while making decisions on model fit.

Measurement invariance was tested by using the nested method (Brown, 2006) which incrementally compares structural invariance (Model 1), weak factorial invariance (Model 2), strong factorial invariance (Model 3) and strict factorial invariance (Model 4) models. Models were compared in the following order Model 1 (structural invariance) and Model 2 (metric invariance), Model 2 and Model 3 (strong factorial invariance) and Model 3 and Model 4 (strict factorial invariance) the results were assessed and a decision was made as to measurement invariance. If there is equal fit to the fit of the compared model or if the fit is worse, it was accepted that limited measurement model was confirmed (Van de Vijver and Leung, 1997).

While measurement models were compared, changes in fit indices were examined and the differences between these indices were noted. First of all, the difference ($\Delta \chi^2$ and $\Delta sd$) between $\chi^2$ statistics and degree of freedom is examined in the models that were compared. Whether the difference obtained as result of comparison is identified by comparing the values in $\chi^2$ Table at $p<.01$ or $p<.05$ levels. If the value in $\chi^2$ Table is bigger than the obtained result, it is accepted that the difference between models is significant (Kline, 2005). In this study, since S-B $\chi^2$ value was taken as $\chi^2$, $T_s$ values were calculated to determine the degree of difference and the significance of the difference between models were identified by comparing the obtained $T_s$ value to the critical value of $p<.05$ included in $\chi^2$ Table. In addition, it is suggested to use fit indices such as $\Delta$CFI, $\Delta$ SRMR, $\Delta$ NFI, $\Delta$GFI, $\Delta$RMSEA and $\Delta$TLI in model comparisons (Cheng and Rensvold, 2002). This study undertook comparisons based on $\Delta$CFI and $\Delta$SRMR values. When the level of acceptance is $n>300$ for these comparisons, the identified values are as follows: $\Delta$CFI$\geq-.010$ and $\Delta$SRMR$\geq.015$ for metric invariance; $\Delta$CFI$\geq-.010$ and $\Delta$SRMR$\leq.010$ for comparing strong and strict factorial invariance. When the level of acceptance is $n<300$; $\Delta$CFI$\leq-.005$ and $\Delta$SRMR$\geq.025$ was identified for metric invariance and $\Delta$CFI$\leq-.005$, $\Delta$SRMR$\leq.005$ in comparisons for strong and strict factorial invariance (Chen, 2007).
3. Findings

3.1. Findings on Measurement Invariance for Gender Groups

The first sub problem in the study, whether PISA 2012 Turkey implementation mathematical literacy provided measurement invariance according to gender groups, was examined under this title. Before presenting the data regarding the measurement invariance between genders, measures of central tendency, coefficients of kurtosis and skewness and KR-20 reliability coefficient were calculated and the results are displayed in Table 5.

<table>
<thead>
<tr>
<th>Gender</th>
<th>N</th>
<th>( \bar{x} )</th>
<th>Mod</th>
<th>S</th>
<th>Range</th>
<th>( K_y )</th>
<th>( B_z )</th>
<th>KR-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>183</td>
<td>4.17</td>
<td>2</td>
<td>2.23</td>
<td>10</td>
<td>.493</td>
<td>-.358</td>
<td>.664</td>
</tr>
<tr>
<td>Male</td>
<td>194</td>
<td>4.23</td>
<td>4</td>
<td>2.37</td>
<td>11</td>
<td>.628</td>
<td>.011</td>
<td>.695</td>
</tr>
</tbody>
</table>

\( K_y \): Coefficient of Skewness  
\( B_z \): Coefficient of Kurtosis  
KR-20: Reliability Coefficient

According to the results presented in Table 5, it can be argued that measures of central tendency for the groups were similar. When coefficients of kurtosis and skewness were examined for normality, the values were found between ±1 which showed the distribution of the data set was close to normal distribution (Rosenthal and Rosnow, 2008). When the reliability coefficients for gender groups were examined, it was observed that reliability was not at the accepted level of 70-.80 range (Nunnally and Bernstein, 1994) but close to level of acceptance at .70. Low number of items closely affects reliability. It is believed that this finding is related to low number of items in the study.

Before Multi-Group Confirmatory Factor Analysis was undertaken, equivalence of covariance matrices for gender groups was tested. Test results are provided in Table 6.

<table>
<thead>
<tr>
<th>Group</th>
<th>( S-B \chi^2(\text{sd}) )</th>
<th>p</th>
<th>( \chi^2/\text{sd} )</th>
<th>RMSEA</th>
<th>GFI</th>
<th>CFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female-Male</td>
<td>41.55 (66)</td>
<td>.99</td>
<td>.62</td>
<td>.00</td>
<td>.98</td>
<td>1.00</td>
<td>.045</td>
</tr>
</tbody>
</table>

Table 6 shows that \( \chi^2/\text{sd} \) ratio was below 2, RMSEA and SRMR values were below .05 and CFI and GFI values were over .95. In this case, it can be argued that there was a high fit between the two covariance matrices for female and male groups.

Results of Multi-Group Confirmatory Factor Analysis, conducted to identify whether CFA results for gender groups and PISA 2012 mathematical literacy provided measurement invariance according to female and male groups, are provided in Table 7.
Table 7: Multi Group Confirmatory Factor Analysis Results of Mathematical Literacy Test Scores for Female and Male Groups

|        | S-B \( \chi^2 \) (sd)* | M.K.** | \( \Delta \chi^2 \) (Asd) | \( \chi^2 \)/sd | \( \Delta \chi^2 \)/Asd | CFI | \( \Delta \) CFI | SRMR | \( \Delta \) SRMR |
|--------|--------------------------|--------|--------------------------|-----------------|-----------------|-----|----------------|------|----------------|-----|
| Female | 34.17 (44)               | -      | -                        | .77             | 1.00            | .044| -              |      | -              |     |
| Male   | 37.89 (44)               | -      | -                        | .86             | 1.00            | .044| -              |      | -              |     |
| Model 1 | 85.79 (110)             | -      | -                        | .78             | 1.00            | .057| -              |      | -              |     |
| Model 2a | 77.75 (99)             | M1-M2  | 8.04 (11)                | .78             | 0               | .047| .01           |      | -              |     |
| Model 3c | 77.75 (99)             | M2-M3  | 0                        | .78             | 0               | .047| 0              |      | -              |     |
| Model 4b | 41.55 (66)             | M3-M4  | 36.2 (33)                | .62             | .16             | 1.00| 0.045         | .002 | -              |     |

*p<.05
**Model comparisons (M=Model)
A. Structural invariance (factor loadings, factor correlations and error variances constant)
B. Weak factorial invariance (factor loadings free, factor correlations and error variances constant)
C. Strong factorial invariance (factor loadings and error variances free, factor correlations constant)
D. Strict factorial invariance (Error variances free, factor loadings and factor correlations constant)

Examination of fit indices related to results of CFA for female and male groups presented in Table 7 shows that \( \chi^2 \)/sd ratio for both groups was below 2, CFI value was over .95 and SRMR value was smaller than .05. In this case, it can be argued that the factorial structure of the test was confirmed for both female and male groups separately.

As can be seen in Table 7, as a result of testing whether factor loadings, factor correlations and error variances in the covariance matrices of the structural model for the gender groups (Model 1) were equal to all groups, it was observed that S-B \( \chi^2 \) and level of freedom ratio were below 2, CFI value was equal to 1 and SRMR value was about .05. In this case, it can be argued that the values for fit indices were at acceptable levels and the model fit was good.

When Model 1 (structural invariance) and Model 2 (weak factorial invariance) where factor loadings were free were compared, it can be argued that \( \Delta \chi^2 \)/Asd ratio and CFI did not change and, \( \Delta \) SRMR was not significant (<.025).

When Model 2 (weak factorial invariance) Model 3 (strong factorial invariance) where factor loadings and error variances were free and factor correlations were constant were compared, it was found that \( \Delta \chi^2 \)/Asd ratio, CFI and SRMR values did not display any changes.

When Model 3 (strong factorial invariance) and Model 4 (strict factorial invariance) where error variances are free and factor loadings and factor correlations were constant were compared; it was found that \( \Delta \chi^2 \)/Asd ratio improved to a small extent. \( T_s \) value for S-B \( \chi^2 \) degree of difference was calculated to identify whether the improvement was significant and the value of 37.63 was obtained. It was observed that the calculated \( T_s \) value was smaller than the critical value included in \( \chi^2 \) distribution table; \( \chi^2_{fark}(33)=47.4, p>.05. \) In this case, it is possible to state that this value was not significant. It was also identified that CFI value did not change and \( \Delta \) SRMR value was not significant (<.025). In this case, it can be argued that the difference between Model 3 and Model 4 was insignificant.
Examination of the results pertaining to all model comparisons shows that Model 1 (structural invariance model) was the best model among the models that were compared to one another. In this case, it can be stated that factor structures were equal for the gender groups on the basis of covariance matrices. Accordingly, it was decided that PISA 2012 Turkey implementation mathematical literacy test provided measurement invariance for female and male groups.

3.2. Findings on Measurement Invariance for Region 1, Region 2 and Region 3
The second sub problem in the study, whether PISA 2012 Turkey implementation mathematical literacy provided measurement invariance according to regions, was examined under this title. Before presenting the data regarding the measurement invariance between regions, measures of central tendency, coefficients of kurtosis and skewness and KR-20 reliability coefficient were calculated and the results are displayed in Table 8.

<table>
<thead>
<tr>
<th>Regions</th>
<th>N</th>
<th>Mode</th>
<th>S</th>
<th>Range</th>
<th>KR-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>141</td>
<td>4.34</td>
<td>2.32</td>
<td>11</td>
<td>.574</td>
</tr>
<tr>
<td>Region 2</td>
<td>109</td>
<td>4.02</td>
<td>2.27</td>
<td>10</td>
<td>.549</td>
</tr>
<tr>
<td>Region 3</td>
<td>127</td>
<td>4.19</td>
<td>2.30</td>
<td>11</td>
<td>.592</td>
</tr>
</tbody>
</table>

$K_r$: Coefficient of Skewness
$B_k$: Coefficient of Kurtosis
KR-20: Reliability Coefficient

According to the results presented in Table 8, it can be argued that measures of central tendency for the groups were similar. When coefficients of kurtosis and skewness were examined for normality, the values were found between ±1 which showed the distribution of the data set was close to normal distribution. When the reliability coefficients for gender groups were examined, it was observed that reliability was close to level of acceptance (.70).

Before Multi-Group Confirmatory Factor Analysis was undertaken, equivalence of covariance matrices for Region 1, Region 2 and Region 3 was tested. Test results are provided in Table 9.

<table>
<thead>
<tr>
<th>Group</th>
<th>$S-B_{x^2}(sd)$</th>
<th>$\chi^2/sd$</th>
<th>RMSEA</th>
<th>GFI</th>
<th>CFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1-Region 2- Region 3</td>
<td>123.98 (132)</td>
<td>.68</td>
<td>.93</td>
<td>.00</td>
<td>.95</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 9 shows that $\chi^2/sd$ ratio was below 2, RMSEA value was 0.0, SRMR value was below .08 and CFI and GFI values were over .95. According to the examination of fit indices, it can be argued that the fit among the three covariance matrices for Region 1, Region 2 and Region 3 groups was high.
Results of Multi-Group Confirmatory Factor Analysis, conducted to identify whether CFA results for gender groups and PISA 2012 mathematical literacy provided measurement invariance according to regions, are provided in Table 10.

<table>
<thead>
<tr>
<th>Region</th>
<th>S-B_{χ^2}(sd)*</th>
<th>M.K.**</th>
<th>Δχ^2/Δsd</th>
<th>χ^2/sd</th>
<th>Δχ^2/Δsd</th>
<th>CFI</th>
<th>ΔCFI</th>
<th>SRMR</th>
<th>Δ SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>30.57 (43)</td>
<td>-</td>
<td>-</td>
<td>0.71</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.046</td>
<td>-</td>
</tr>
<tr>
<td>Region 2</td>
<td>44.97 (44)</td>
<td>-</td>
<td>-</td>
<td>1.02</td>
<td>-</td>
<td>0.93</td>
<td>-</td>
<td>0.067</td>
<td>-</td>
</tr>
<tr>
<td>Region 3</td>
<td>32 (43)</td>
<td>-</td>
<td>-</td>
<td>0.74</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.05</td>
<td>-</td>
</tr>
<tr>
<td>Model 1^4</td>
<td>173.26 (176)</td>
<td>-</td>
<td>-</td>
<td>0.98</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.069</td>
<td>-</td>
</tr>
<tr>
<td>Model 2^#</td>
<td>144.95 (154)</td>
<td>M1-M2</td>
<td>28.31 (22)</td>
<td>0.94</td>
<td>0.04</td>
<td>1.00</td>
<td>0</td>
<td>0.060</td>
<td>0.009</td>
</tr>
<tr>
<td>Model 3^C</td>
<td>129.15 (132)</td>
<td>M2-M3</td>
<td>15.8 (22)</td>
<td>0.97</td>
<td>0.03</td>
<td>1.00</td>
<td>0</td>
<td>0.056</td>
<td>0.004</td>
</tr>
<tr>
<td>Model 4^D</td>
<td>161.80 (154)</td>
<td>M3-M4</td>
<td>-32.65 (-22)</td>
<td>1.05</td>
<td>-0.08</td>
<td>0.99</td>
<td>0.01</td>
<td>0.067</td>
<td>-0.11</td>
</tr>
</tbody>
</table>

*p<.05

**Model comparisons (M=Model)

A Structural invariance (factor loadings, factor correlations and error variances constant)
B Weak factorial invariance (factor loadings free, factor correlations and error variances constant)
C Strong factorial invariance (factor loadings and error variances free, factor correlations constant)
D Strict factorial invariance (Error variances free, factor loadings and factor correlations constant)

Examination of fit indices related to results of CFA for regions presented in Table 10 shows that \( \chi^2/\text{sd} \) ratio for all three groups was below 2, CFI value was over .95 for Region 1 and Region 3 and .93 for Region 2 and SRMR value for all regions is smaller than .08. In this case, it can be argued that the factorial structure of the test was confirmed for all regions separately.

As can be seen in Table 10, as a result of testing whether factor loadings, factor correlations and error variances in the covariance matrices of the structural model for the regions (Model 1) were equal to all groups, it was observed that S-B_{χ^2} and level of freedom ratio were below 2, CFI value was over .95 and SRMR value was smaller than .08. In this case, it can be argued that the values for fit indices were at acceptable levels and the model fit was good.

When Model 1 (structural invariance) and Model 2 (weak factorial invariance) where factor loadings were free were compared, \( \Delta \chi^2/\Delta \text{sd} \) ratio seemed improved. \( T_s \) value which was used to determine whether the improvement was significant was calculated to be 28.126. It was found that this value was smaller than the critical value presented in \( \chi^2 \) distribution Table and the value was not significant, \( \chi^2_{\text{fark}}(22)=33.924, p>.05 \). It was also found that CFI did not change and, ΔSRMR was smaller than .025. In this case, it can be stated that a small amount of improvement was observed but the difference between Model 1 and Model 2 was not significant.

When Model 2 (weak factorial invariance) Model 3 (strong factorial invariance) where factor loadings and error variances were free and factor correlations were constant were compared, it was found that \( \Delta \chi^2/\Delta \text{sd} \) ratio was found to be improved. \( T_s \) value which was used to determine whether S-B_{χ^2} degree of difference was significant was calculated to be 0.772. It was found that \( T_s \) value was smaller than the critical value...
presented in $\chi^2$ distribution Table and the value was not significant, $\chi^2_{fark}(22)=33.924$, p>.05. While there was no difference in CFI value, $\Delta$SRMR value was not significant. In this case, it is possible to state that no significant differences exist between weak factorial invariance and strong factorial invariance.

When Model 3 (strong factorial invariance) and Model 4 (strict factorial invariance) where error variances are free and factor loadings and factor correlations were constant were compared; it was seen that the fit according to $\Delta\chi^2/\Delta sd$ ratio, $\Delta$CFI and $\Delta$SRMR values was not good.

Examination of the results pertaining to all model comparisons shows that Model 1 (structural invariance model) was the best model among the models that were compared to one another. In this case, it can be stated that factor structures were equal for the region groups on the basis of covariance matrices. Accordingly, it was decided that PISA 2012 Turkey implementation mathematical literacy test provided measurement invariance for Region 1, Region 2 and Region 3.

### 3.3. Findings on Measurement Invariance for OECD Member Countries

The third sub problem in the study, whether PISA 2012 Turkey implementation mathematical literacy provided measurement invariance according to OECD member countries, was examined under this title. Before presenting the data regarding the measurement invariance among OECD member countries, measures of central tendency, coefficients of kurtosis and skewness and KR-20 reliability coefficient were calculated for these countries’ mathematical literacy test score distribution and the results are displayed in Table 11.

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>$\bar{x}$</th>
<th>Mode</th>
<th>S</th>
<th>Range</th>
<th>$K_r$</th>
<th>$B_s$</th>
<th>KR-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1198</td>
<td>4.99</td>
<td>4</td>
<td>2.35</td>
<td>11</td>
<td>.206</td>
<td>-.506</td>
<td>.70</td>
</tr>
<tr>
<td>Austria</td>
<td>373</td>
<td>5.18</td>
<td>4</td>
<td>2.22</td>
<td>11</td>
<td>.138</td>
<td>.252</td>
<td>.68</td>
</tr>
<tr>
<td>Belgium</td>
<td>661</td>
<td>5.65</td>
<td>6</td>
<td>2.44</td>
<td>11</td>
<td>-.061</td>
<td>-.434</td>
<td>.72</td>
</tr>
<tr>
<td>Canada</td>
<td>1679</td>
<td>5.20</td>
<td>5</td>
<td>2.26</td>
<td>11</td>
<td>.097</td>
<td>-.388</td>
<td>.67</td>
</tr>
<tr>
<td>Chile</td>
<td>532</td>
<td>4.2</td>
<td>4</td>
<td>2.15</td>
<td>10</td>
<td>.217</td>
<td>-.548</td>
<td>.64</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>429</td>
<td>5.85</td>
<td>5</td>
<td>2.27</td>
<td>11</td>
<td>-.025</td>
<td>.235</td>
<td>.68</td>
</tr>
<tr>
<td>Denmark</td>
<td>558</td>
<td>5.07</td>
<td>6</td>
<td>2.40</td>
<td>10</td>
<td>-.038</td>
<td>-.686</td>
<td>.73</td>
</tr>
<tr>
<td>Estonia</td>
<td>404</td>
<td>5.41</td>
<td>5</td>
<td>2.11</td>
<td>11</td>
<td>.174</td>
<td>-.074</td>
<td>.63</td>
</tr>
<tr>
<td>Finland</td>
<td>665</td>
<td>5.26</td>
<td>5</td>
<td>2.27</td>
<td>11</td>
<td>.075</td>
<td>-.334</td>
<td>.69</td>
</tr>
<tr>
<td>France</td>
<td>353</td>
<td>5.13</td>
<td>4</td>
<td>2.36</td>
<td>11</td>
<td>.203</td>
<td>-.415</td>
<td>.71</td>
</tr>
<tr>
<td>Germany</td>
<td>387</td>
<td>5.39</td>
<td>5</td>
<td>2.32</td>
<td>11</td>
<td>.068</td>
<td>-.517</td>
<td>.71</td>
</tr>
<tr>
<td>Greece</td>
<td>391</td>
<td>4.03</td>
<td>3</td>
<td>2.25</td>
<td>11</td>
<td>.520</td>
<td>-.007</td>
<td>.70</td>
</tr>
<tr>
<td>Hungary</td>
<td>361</td>
<td>4.91</td>
<td>4</td>
<td>2.33</td>
<td>11</td>
<td>.227</td>
<td>-.331</td>
<td>.70</td>
</tr>
<tr>
<td>Iceland</td>
<td>258</td>
<td>5.28</td>
<td>5</td>
<td>2.39</td>
<td>10</td>
<td>-.134</td>
<td>-.558</td>
<td>.71</td>
</tr>
<tr>
<td>Ireland</td>
<td>376</td>
<td>5.22</td>
<td>6</td>
<td>2.11</td>
<td>11</td>
<td>-.057</td>
<td>-.158</td>
<td>.63</td>
</tr>
<tr>
<td>Israel</td>
<td>561</td>
<td>4.46</td>
<td>4</td>
<td>2.49</td>
<td>10</td>
<td>.209</td>
<td>-.648</td>
<td>.74</td>
</tr>
<tr>
<td>Italy</td>
<td>2643</td>
<td>5.10</td>
<td>5</td>
<td>2.40</td>
<td>11</td>
<td>.155</td>
<td>-.500</td>
<td>.71</td>
</tr>
<tr>
<td>Japan</td>
<td>477</td>
<td>5.70</td>
<td>7</td>
<td>2.43</td>
<td>11</td>
<td>-.217</td>
<td>-.666</td>
<td>.70</td>
</tr>
<tr>
<td>Korea</td>
<td>394</td>
<td>6.04</td>
<td>5</td>
<td>2.57</td>
<td>11</td>
<td>-.037</td>
<td>-.769</td>
<td>.73</td>
</tr>
</tbody>
</table>
Before Multi-Group Confirmatory Factor Analysis was undertaken, equivalence of covariance matrices for OECD countries was tested. Test results are provided in Table 12.

Table 12: Equivalence of Covariance Matrices Test Results of Mathematical Literacy Test Scores for OECD countries

<table>
<thead>
<tr>
<th>Grup</th>
<th>S-B $\chi^2$(sd)</th>
<th>$p$</th>
<th>$\chi^2$/sd</th>
<th>RMSEA</th>
<th>GFI</th>
<th>CFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>OECD Üyesi Ülkeler</td>
<td>8317.72 (2178)</td>
<td>.00</td>
<td>3.81</td>
<td>.064</td>
<td>.94</td>
<td>.85</td>
<td>.092</td>
</tr>
</tbody>
</table>

As seen in Table 12, the ratio of the $\chi^2$/sd is lower than 0.05, the value of SRMR is higher than 0.08, and the value of CFI is higher than 0.90. According to the fit indices, it could be said that the covariance matrices of the countries are consistent in a middle level.

Because of so many CFA results of the countries, the MGCFA results are separated and presented in Table 13.

Table 13: Confirmatory Factor Analysis Results of Mathematical Literacy Test Scores for OECD Countries

<table>
<thead>
<tr>
<th>Country</th>
<th>S-B $\chi^2$(sd)</th>
<th>$\chi^2$/sd</th>
<th>CFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>175.34 (44)</td>
<td>3.98</td>
<td>.95</td>
<td>.037</td>
</tr>
<tr>
<td>Austria</td>
<td>77.42 (44)</td>
<td>1.75</td>
<td>.95</td>
<td>.047</td>
</tr>
<tr>
<td>Belgium</td>
<td>116.13 (44)</td>
<td>2.63</td>
<td>.96</td>
<td>.041</td>
</tr>
<tr>
<td>Canada</td>
<td>129.03 (44)</td>
<td>2.93</td>
<td>.97</td>
<td>.028</td>
</tr>
<tr>
<td>Chile</td>
<td>66.81 (44)</td>
<td>1.51</td>
<td>.96</td>
<td>.039</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>63.11 (44)</td>
<td>1.43</td>
<td>.97</td>
<td>.038</td>
</tr>
<tr>
<td>Denmark</td>
<td>82.81 (44)</td>
<td>1.88</td>
<td>.97</td>
<td>.038</td>
</tr>
<tr>
<td>Estonia</td>
<td>74.75 (44)</td>
<td>1.69</td>
<td>.93</td>
<td>.046</td>
</tr>
<tr>
<td>Finland</td>
<td>157.97 (44)</td>
<td>3.59</td>
<td>.92</td>
<td>.048</td>
</tr>
<tr>
<td>France</td>
<td>81.44 (44)</td>
<td>1.85</td>
<td>.96</td>
<td>.045</td>
</tr>
<tr>
<td>Germany</td>
<td>72.39 (44)</td>
<td>1.64</td>
<td>.96</td>
<td>.042</td>
</tr>
<tr>
<td>Greece</td>
<td>95.02 (44)</td>
<td>2.15</td>
<td>.94</td>
<td>.048</td>
</tr>
<tr>
<td>Hungary</td>
<td>107.32 (44)</td>
<td>2.43</td>
<td>.93</td>
<td>.053</td>
</tr>
</tbody>
</table>
According to the results presented in Table 13, $\chi^2$/sd ratios for 29 countries were found to be below 3 and $\chi^2$/sd ratios for 3 countries were found to be below 5. It was identified that $\chi^2$/sd ratios for Spain and Italy were over 5. However, it is believed that largeness of sample sizes in these two countries affected the ratio and therefore other fit indices were examined. CFI values for all countries were found to be higher than .90. It was identified that SRMR values for all countries other than Hungary were below .05 while the SRMR value for Hungary was 0.053. General examination of fit indices shows that models for all countries were confirmed for 34 OECD member countries.

Results of Multi-Group Confirmatory Factor Analysis, conducted to identify whether PISA 2012 mathematical literacy provided measurement invariance among OECD countries, are provided in Table 14.

### Table 14: Multi Group Confirmatory Factor Analysis Results of Mathematical Literacy Test Scores for OECD Member Countries

<table>
<thead>
<tr>
<th>Country</th>
<th>$S^2/B_{XX}$(sd)</th>
<th>M.K.</th>
<th>$\Delta \chi^2$(Asd)</th>
<th>$\chi^2$/sd</th>
<th>$\Delta \chi^2$/Asd</th>
<th>CFI</th>
<th>ACFI</th>
<th>SRMR</th>
<th>A SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>50.77 (44)</td>
<td>1.15</td>
<td>.99</td>
<td>.044</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td>61.03 (44)</td>
<td>1.38</td>
<td>.96</td>
<td>.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td>103.80 (44)</td>
<td>2.35</td>
<td>.96</td>
<td>.041</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td>288.72 (44)</td>
<td>6.56</td>
<td>.96</td>
<td>.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>62.08 (44)</td>
<td>1.41</td>
<td>.98</td>
<td>.036</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korea</td>
<td>77.97 (44)</td>
<td>1.77</td>
<td>.97</td>
<td>.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luxembourg</td>
<td>101.01 (44)</td>
<td>2.29</td>
<td>.96</td>
<td>.048</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>211.52 (44)</td>
<td>4.80</td>
<td>.91</td>
<td>.032</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holland</td>
<td>72.38 (44)</td>
<td>1.64</td>
<td>.96</td>
<td>.046</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Zealand</td>
<td>73.35 (44)</td>
<td>1.66</td>
<td>.95</td>
<td>.045</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>63.34 (44)</td>
<td>1.43</td>
<td>.97</td>
<td>.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td>66.21 (44)</td>
<td>1.50</td>
<td>.98</td>
<td>.041</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portugal</td>
<td>90.31 (44)</td>
<td>2.05</td>
<td>.95</td>
<td>.046</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovakia</td>
<td>68.93 (44)</td>
<td>1.56</td>
<td>.98</td>
<td>.039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td>97.27 (44)</td>
<td>2.21</td>
<td>.95</td>
<td>.044</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>235.04 (44)</td>
<td>5.34</td>
<td>.95</td>
<td>.035</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>38.70 (44)</td>
<td>0.87</td>
<td>1.00</td>
<td>.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>93.33 (44)</td>
<td>2.12</td>
<td>.97</td>
<td>.033</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>41.60 (44)</td>
<td>0.94</td>
<td>1.00</td>
<td>.033</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>130.18 (44)</td>
<td>2.95</td>
<td>.95</td>
<td>.038</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States of America</td>
<td>73.37 (44)</td>
<td>1.66</td>
<td>.96</td>
<td>.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p<.05  
**Model comparisons (M=Model)  
A Structural invariance (factor loadings,factor correlations and error variances constant)
As can be seen in Table 14, as a result of testing whether factor loadings, factor correlations and error variances in the covariance matrices of the structural model for the regions (Model 1) were equal to all groups, it was observed that $S$-value of $\chi^2$ and level of freedom ratio were below 5, CFI value was below .90 and SRMR value was smaller than .1. In this case, it can be argued that the values for fit indices were not at acceptable levels and the model was not confirmed. In this case, it is possible that structural invariance was not provided.

When Model 1 (structural invariance) and Model 2 (weak factorial invariance) where factor loadings were free were compared, it can be argued that $\Delta \chi^2/\Delta sd$ ratio became worse. Also, $\Delta$CFI and $\Delta$SRMR values did not change. In this case, it can be argued that the difference between Model 1 and Model 2 was not significant.

When Model 2 (weak factorial invariance) Model 3 (strong factorial invariance) where factor loadings and error variances were free and factor correlations were constant were compared, it was found that $\Delta \chi^2/\Delta sd$ ratio improved significantly. Value for $S$-degree of difference was calculated to identify whether the improvement was significant and the value of 4910.97 was obtained. It was observed that the calculated $T_s$ value was bigger than the critical value included in $\chi^2$ distribution table; $\chi^2_{fark}(363)=394.626$, $p>.05$. In this case, it is possible to state that this value was significant. $\Delta$CFI value -0.12 ($<-0.01)$ and $\Delta$SRMR value ($\geq .025$) were also identified to be significant. In this case, it can be reported that the difference between weak factorial invariance and strong factorial invariance was significant. It can be stated that measurement invariance was not provided since strong factorial invariance model had a good fit.

When Model 3 (strong factorial invariance) and Model 4 (strict factorial invariance) where error variances are free and factor loadings and factor correlations were constant were compared; it was seen that $\Delta \chi^2/\Delta sd$ ratio, $\Delta$CFI and $\Delta$SRMR values did not change. In this case, it can be argued that the difference between Model 3 and Model 4 was insignificant.

Examination of the findings regarding all model comparisons shows that Model 4 (strong factorial invariance) was the best model among the 4 models that were compared to one another. Accordingly, it was decided that PISA 2012 Turkey implementation mathematical literacy test was not equal for OECD member countries and did not provide measurement invariance.

4. Results and Discussion

Results on the measurement invariance of PISA 2012 mathematical literacy based on OECD member countries and gender and regions in Turkey are as follows:
In order to determine whether PISA 2012 mathematical literacy provided measurement invariance for gender groups, test statistics, normality and reliability calculations were undertaken as the first step. As a result of these calculations, it was found that the data set presented a distribution close to normal. As a result of reliability calculations, it was found KR-20 internal consistency reliability coefficient did not display high values due to lower number of items but it was close to acceptable levels.

A medium level fit was identified between the covariance matrices of the gender groups based on the Equivalence of Covariance Matrices Test for gender groups. Measurement invariance of PISA 2012 mathematical literacy according to gender groups was tested via Multi Group Confirmatory Factor Analysis based on four models. Model 1, the basic model, was established with the hypothesis that factorial structures were equal. Fit indices for Model 1 were found to be at acceptable levels. Based on the comparison of Model 2, Model 3 and Model 4—established as alternatives to Model 1—according to nested model, it was found that fit became worse compared to Model 1. Model 1, formed with the hypothesis that factorial structures for gender groups were equal, was identified to be the best working model. In the framework of the findings obtained in the study, it was found that PISA 2012 mathematical literacy provided measurement invariance between the gender groups in Turkey.

In order to determine whether PISA 2012 mathematical literacy provided measurement invariance according to regions in Turkey (Region 1, Region 2, Region 3), test statistics, normality and reliability calculations were undertaken. As a result of these calculations, it was found that the data set presented a distribution close to normal. As a result of reliability calculations, it was found KR-20 internal consistency reliability coefficient was close to acceptable levels.

A high level fit was identified between the covariance matrices for Region 1, Region and Region 3 based on the Equivalence of Covariance Matrices Test for regions. Measurement invariance of PISA 2012 mathematical literacy according to regions was tested via Multi Group Confirmatory Factor Analysis based on four models. Model 1, the basic model, was established with the hypothesis that factorial structures were equal. Fit indices for Model 1 were found to be at acceptable levels. Based on the comparison of Model 2, Model 3 and Model 4—established as alternatives to Model 1—according to nested model, it was found that fit became worse compared to Model 1. Model 1, formed with the hypothesis that factorial structures for Region 1, Region 2 and Region 3 were equal, was identified to be the best working model. In the framework of the findings obtained in the study, it was found that PISA 2012 mathematical literacy provided measurement invariance according to regions in Turkey.

In order to determine whether PISA 2012 mathematical literacy provided measurement invariance among OECD member countries, test statistics, normality and reliability calculations were undertaken. As a result of these calculations, it was accepted that the data set presented a distribution close to normal. As a result of reliability calculations, it was found KR-20 internal consistency reliability coefficient was below acceptable levels for some countries while it was at acceptable level for some others.
The covariance matrices among OECD countries based on the Equivalence of Covariance Matrices Test for countries showed that the model for the covariance matrices for countries did not provide good fit. Measurement invariance of PISA 2012 mathematical literacy according to OECD countries was tested via Multi Group Confirmatory Factor Analysis based on four models. Model 1, the basic model, which established with the hypothesis that factorial structures were equal, was not found to be at acceptable levels in terms of its fit indices. Therefore the model was not confirmed. In this case, it can be claimed that structural invariance was not provided according to OECD member countries. Based on the comparison of Model 2, Model 3 and Model 4-established as alternatives to Model 1- according to nested model, it was found that fit for Model 3 significantly improved and strong factorial invariance was confirmed. It was identified that the model that worked best for PISA 2012 mathematical literacy for OECD member countries was Model 3 (strong factorial invariance) and measurement invariance was only provided at strong factorial invariance level. Findings of this study which assessed measurement invariance for PISA 2012 Mathematical literacy through booklet 8 show that measurement invariance was provided for gender and regions in Turkey but not among OECD member countries.

Findings obtained in the scope of gender variable are parallel to Kıbrıslıoğlu’s (2015) study that pointed to provision of measurement invariance for PISA 2012 mathematical learning model based on gender in Turkey, China-Shanghai and Indonesia. However, Öğretmen’s (2010) study on the science achievements of students who participated in 1999 TIMSS-R reported that only the model related to metric invariance was confirmed according to gender and Uyar and Doğan’s (2014) study on PISA 2009 found that learning strategies model in Turkey sample confirmed structural and metric invariance models according to gender.

When evaluated based on the variable of regions, the findings of the study are parallel to Bahadır’s (2012) study that reported measurement invariance among geographical regions for PISA 2009 reading skills model and Uyar and Doğan’s (2014) study that reported measurement invariance among statistical regions for PISA 2009 learning strategies model in Turkey implementation.

When the findings of the study are evaluated based on the variable of OECD member countries, it can be claimed that adapting the measurement tools that are used in wide scale exams to different languages and cultures creates limitations. Some studies reported that measurement invariance was not provided in the measurement tools used by individuals with different languages and cultures (Ercikan and Koh 2005; Öğretmen, 2006; Wu et.al., 2007; et.al., 2013; Kıbrıslıoğlu, 2015; Karakoğ Alatlı, 2016). In addition, some wide scale studies conducted in various countries reported that measurement invariance was provided among different countries (Marsh et. al, 2006; Akyıldız, 2009). However, the inability to provide measurement invariance should be taken into consideration while assessing the results of wide scale exams and while making comparisons especially among countries and reporting relevant interpretations.

When all results are evaluated, it was identified that findings obtained for PISA 2012 Mathematical literacy in Turkey implementation were comparable and
interpretable based on gender and regions. However, comparability of results obtained from participants from different cultures who speak different languages and relevant comments related to this topic are open to discussion due to the inability to provide measurement invariance which points to the fact that the implemented measurement tool did not mean the same things for all participants. It should be remembered that measurement invariance differs according to the measured characteristic or the type of implementation in wide scale assessment and measurement practices and that measurement invariance studies should be conducted before making comparisons and providing comments based on the obtained results.

References


