




**TIMSS Dördüncü Sınıf Matematik Testinin OECD Üyesi Ülkelere
Göre Ölçme Değişmezliğinin İncelenmesi**
**An Investigation of Measurement Invariance of the TIMSS Fourth-Grade
Mathematics Assessment Across OECD Member Countries**



Yazar Bilgisi/ Author Information

Ayşenur TAVLICA

 Millî Eğitim Bakanlığı, İstanbul/Türkiye, aysenurtavlica@gmail.com

 **Güçlü ŞEKERCİOĞLU**

Doç.Dr., Eğitim Fakültesi/Akdeniz Üniversitesi, Antalya/Türkiye, guclus@akdeniz.edu.tr

Makale Bilgisi/ Article Info

Makale Türü/ Article Type : Araştırma Makalesi / Research Article
Geliş Tarihi/ Received : 06.03.2025
Kabul Tarihi /Accepted : 28.03.2025
Yayın Tarihi/Published : 27.06.2025

Atıf / Cite

Tavlica, A. & Şekercioğlu, G. (2025). TIMSS dördüncü sınıf matematik testinin OECD üyesi ülkelere göre ölçme değişmezliğinin incelenmesi. *EDUCATIONE*, 4(1), 154-172.

Bu araştırmada, TIMSS 2015 uygulaması çerçevesinde uygulanan dördüncü sınıf matematik başarı testinden elde edilen puanların faktör deseninin ölçme değişmezliğinin OECD üyesi ülkelere göre incelenmesi amaçlanmıştır. TIMSS 2015 uygulamasına 24 ülkeden toplam 132.226 öğrenci katılmıştır. İlişkisel tarama modelindeki bu araştırma rastlantısal olarak seçilen, 7. Kitapçığı alan 9.641 öğrenci verisi üzerinden yürütülmüştür. Araştırma kapsamında, ölçme değişmezliği analizlerine geçmeden önce ölçme değişmezliği analizlerinin temel sayıtları olan normallik test edilmiş; güvenilirlik katsayıları belirlenmiş, ölçme değişmezliği test edilecek tüm gruplar için doğrulayıcı faktör analizi yapılmış ve kovaryans matrislerinin eşitliği test edilmiştir. Yapılan analizler sonucunda sayıtların doğrulandığı tespit edilmiştir. Daha sonra çoklu-grup doğrulayıcı faktör analizi ile ölçme değişmezliği analizleri gerçekleştirilmiştir. Analizler tek faktörlü desen üzerinden gerçekleştirilmiştir. Araştırma sonucundan, ölçme değişmezliği modelleri arasında en iyi çalışan modelin güçlü faktöriyel değişmezlik modeli olduğu belirlenmiştir. Bu bulgudan yola çıkılarak TIMSS 2015 dördüncü sınıf matematik başarı testi puanlarının OECD üyesi ülkeler arasında ölçme değişmezliğini sağlamadığı, diğer bir ifadeyle faktör deseninin karşılaştırılan ülkeler için aynı olmadığı ve dolayısıyla söz konusu ülkeler arasında karşılaştırma yapmaya ilişkin şüpheleri arttırdığı sonucuna ulaşılmıştır.

Anahtar Kelimeler: *Ölçme değişmezliği, matematik başarı, TIMSS, OECD*

Abstract

This study aimed to examine the measurement invariance of the factor structure of the scores obtained from the fourth-grade mathematics achievement test administered within the framework of TIMSS 2015 across OECD member countries. A total of 132226 students from 24 countries participated in TIMSS 2015. This correlational survey model research was conducted on the data of 9641 randomly selected students who took the 7th booklet. Within the scope of the study, normality was tested before the measurement invariance analyses. Normality is a basic assumption of measurement invariance analyses. Reliability coefficients were determined, and confirmatory factor analysis was performed for all groups to be tested for measurement invariance. Lastly, the equality of covariance matrices was tested. As a result of the analyses, it was determined that the assumptions were confirmed. Then, multi-group confirmatory factor analysis and measurement invariance analyses were performed on a single-factor design. The research determined that the best working model among the measurement invariance models was the strong factorial invariance model. Based on this finding, it was concluded that TIMSS 2015 fourth-grade mathematics achievement test scores did not provide measurement invariance among OECD member countries; in other words, the factorial design was not the same for the countries compared, thus raising doubts about making comparisons between these countries.

Anahtar Kelimeler: *Measurement invariance, mathematics achievement, TIMSS, OECD*



INTRODUCTION

Educational goals have undergone significant changes, especially since the 1950s. Science and technology, which are among the most important factors determining a country's status, have been accepted as the starting point of social progress (Bayır, Çakıcı, & Atalay, 2016). Information technologies, which have left their mark on the current era, have increased the importance and resources allocated to the education of the positive sciences in many countries worldwide, creating a competitive environment, particularly among nations that rapidly develop their technologies. Countries that have renewed their curricula in this direction have particularly focused on science and mathematics education (Çepni, Ayas, Johnson, & Turgut, 1997). This trend continues to grow today. Human resources play a crucial role in information societies, as they are central to the development of countries, improving living standards, quality, and efficiency (Ergül, 1999). Therefore, this situation is closely related to education policies, as they aim to equip the workforce with tools appropriate to the demands of the age in information societies. Education and training programs must be continually updated in line with the continuous development of technology and evolving information. For this reason, several important international organizations conduct large-scale research in education.

With the development of technology, different areas of competition emerge in countries trying to meet the demands of the modern era. Global competition, particularly in education, necessitates the constant review of the education systems that countries implement or plan to implement (Ministry of National Education [MEB], 2014). In this context, evaluations conducted at the basic education level are believed to contribute to more accurate future predictions. Consequently, emphasis is placed on measurement and evaluation studies for students receiving education at this level.

Many countries conduct research to critically evaluate their education systems to identify deficiencies and address these points (MEB, 2010). Evaluators use the results of these surveys to compare student performance and assess the effectiveness of educational policies and practices in participating countries (Gierl, 2000). One such survey in which Turkey participates is the Trends in International Mathematics and Science Study (TIMSS). TIMSS is designed to assess students' knowledge and skills in mathematics and science and is a project of the Netherlands-based International Association for the Evaluation of Educational Achievement (IEA). The IEA has been conducting international comparative studies on student educational achievement since 1959, pioneering research on international achievement related to various

methods used in education and learning across countries, thereby contributing to mutual learning about effective educational approaches. TIMSS is administered to students in the fourth and eighth grades at four-year intervals. In each cycle, schools and classes are randomly selected to reflect the country as a whole. Countries can participate in the test at the fourth-grade level only, at the eighth-grade level only, or at both levels. Additionally, in countries where TIMSS may be too challenging for fourth-grade students, the assessment can be administered to fifth or sixth-grade students, or ninth-grade students instead of eighth-grade students (TIMSS, 2015).

The overall aim of TIMSS is to measure the achievement of fourth and eighth-grade students in mathematics and science in the participating countries. It also aims to determine and evaluate how education and training occur in schools, assess the effectiveness and efficiency of the education system, and analyze inter-country differences in education systems (TIMSS, 2015). Another aim of TIMSS is to provide important background information that can be used to improve teaching and learning in mathematics and science. To achieve this, achievement tests and various questionnaires are used to collect information on students' performance in science and mathematics, the education systems, curricula, student characteristics, and teacher and school characteristics.

TIMSS monitors changes in student achievement at regular intervals and investigates whether new or revised educational policies affect achievement. The results of TIMSS and PISA (Program for International Student Assessment), which were published after their initial assessments, led to the initiation of radical changes in the education systems of participating countries. Following the first TIMSS results announced in the UK in 1997, governmental organizations began analyzing TIMSS data to assess students' strengths and weaknesses. Switzerland's regular efforts to test student performance before participating in PISA were solidified only after the first PISA results were announced. Similarly, Germany's first PISA results, announced in 1995, were found to be below the expected level among teachers, scientists, and politicians. As a result, politicians decided to reform the German education system, which had not used standardized tests prior to the PISA implementation, to align it with national education standards (Rutkowski, von Davier, & Rutkowski, 2013).

The development of the items to be included in the TIMSS achievement tests and the process of inclusion in the test is coordinated by experts at the TIMSS&PIRLS Study Centre, headquartered at Boston University. The mathematics and science questions in TIMSS are prepared jointly by country representatives within the framework of



predetermined outcomes. The prepared questions are examined by the IEA's science and mathematics item review committee and scoring keys are prepared for open-ended questions. Substitute and main questions are then analyzed in draft blocks and the questions are finalized. The prepared questions are tested with the pilot application after translation and adaptation processes in the participating countries. The questions that are sufficient in terms of psychometric properties are combined with the previous application questions and included in the final application one year after the pilot application. Approximately half of the items in TIMSS 2015 consist of multi-choice questions and half of them consist of long and short answer questions. At both grade levels (fourth and eighth grade), science and math items consist of 28 blocks. Fourteen of these blocks were science blocks and 14 were math blocks. These blocks were distributed to 14 test booklets in blocks of four, two for science and two for mathematics. One of the two blocks in science and one of the two blocks in mathematics is common between the two booklets for test equating between the forms (TIMSS, 2015).

Theoretical Basis of Research

Measurement Invariance

In research in the fields of education and psychology, groups are usually compared with psychological traits. A measurement instrument whose psychometric properties are considered adequate in one cultural group is adapted and applied to another cultural group. This is one of the main approaches adopted in cross-cultural studies. Researchers generally assume that the measurement instrument measures the same construct in all cultural groups, but this assumption needs to be tested (Milfont & Fischer, 2010). The validity of cross-cultural comparisons is vital for many applications in educational and psychological research. The validity evidence of cross-cultural research is that test scores obtained from different countries measure the same construct (Wu, Li, & Zumbo, 2007). TIMSS is an application that evaluates and compares countries' science and mathematics achievements with a single measurement instrument. To discuss the significance of the comparisons and score analyses, it is necessary to verify that the measurement instrument measures the same traits in different groups (Uzun & Öğretmen, 2010). Ensuring measurement invariance between groups is a logical requirement for making cross-group comparisons, but measurement invariance has rarely been tested in organizational research (Vandenberg & Lance, 2000). According to Mark and Wan (2005), inferences made in

cases where measurement invariance is not proven to exist are not scientific, and it is not possible to interpret the differences between groups.

Research on measurement invariance entered the literature more than 50 years ago, but since statistical techniques for testing invariance have become available only recently, researchers are now more frequently expected to test measurement invariance today (Putnick & Bornstein, 2016). Jöreskog (1971) was the first researcher to write about the equality of factor structures. The concept of measurement invariance was introduced and later tested by Shavelson and Muthe'n (1989). are now more frequently expected to test measurement invariance today

Measurement invariance is a crucial psychometric property that ensures scales accurately measure latent variables. In other words, constructs associated with observed variables should be consistent across different groups, such as countries, cultural groups, time periods, or regions within countries. When the same construct is measured consistently across these diverse contexts, researchers conclude that the measurement is invariant for those groups (Horn & McArdle, 1992; Meredith, 1993; Vandenberg & Lance, 2000; Şekercioğlu, 2018). Ensuring measurement invariance validates intergroup comparisons based on the latent variable; simultaneously, it allows researchers to compare and evaluate the formation, determinants, and outcomes of latent variable scores in future studies (Schoot, Lugtig, & Hox, 2012; Cheung & Rensvold, 2002). According to Horn and McArdle (1992), measurement invariance is essential for making accurate inferences and interpretations in studies involving variables such as age, gender, or culture. The fundamental question of measurement invariance is whether the quality measured yields the same results when observing phenomena under different conditions. Without evidence of invariance, making scientific inferences about differences between individuals and groups becomes impossible

Measurement invariance is usually revealed by testing four hierarchical models with multi-group confirmatory factor analysis. These models are named as configural invariance (baseline model), weak factorial invariance (metric invariance), strong factorial invariance (scalar invariance), and strict factorial invariance (residual variance invariance). These four models with progressivity should be tested sequentially (Meredith, 1993). Tests of invariance from weak to strong techniques are demonstrated using multi-group confirmatory factor analyses and structural equation modeling (Horn & McArdle, 1992).



This research's problem is to analyze the state of measurement invariance in 24 OECD countries using TIMSS 2015 data. Accordingly, measurement invariance analyses were conducted using the four models mentioned above.

METHOD

Research Design

This study aims to determine whether the items of the 4th-grade mathematics achievement test, included as a cognitive test in TIMSS 2015, are equivalent across OECD countries. To achieve this, the study was conducted using a correlation survey model, which is appropriate for examining the relationships between the achievement test items utilized in TIMSS 2015 among different groups. In scientific research, the relational survey model is employed to assess the level and direction of relationships between two or more variables (Karasar, 2005).

Research Data

TIMSS 2015 was administered to a total of 253.546 students from 45 countries at the 4th-grade level. Of these, 132.226 students from 24 OECD member countries participated in the research. Specifically, 9.461 students from this group participated in the administration of the 7th booklet, which forms the basis of the research data. The data for this study were obtained from the official TIMSS application website (www.timssandpirls.bc.edu). The data are available to all researchers and no authorization is required. The distribution of students by country is presented in Table 1., which outlines the countries included in the research.

Table 1.

Sample Distribution by Country

| Country | n | % |
|--------------------------|-----|------|
| Germany | 285 | 3.01 |
| United States of America | 720 | 7.61 |
| Australia | 431 | 4.56 |
| Belgium | 384 | 4.06 |
| Czech Republic | 376 | 3.97 |
| Finland | 357 | 3.77 |
| Holland | 321 | 3.39 |
| England | 285 | 3.01 |
| Ireland | 310 | 3.28 |
| Spain | 538 | 5.69 |

| | | |
|-------------|-----|------|
| Sweden | 284 | 3.00 |
| Italy | 307 | 3.24 |
| Japan | 314 | 3.32 |
| Canada | 893 | 9.44 |
| Korea | 331 | 3.50 |
| Hungary | 354 | 3.74 |
| Norway | 286 | 3.02 |
| Polonia | 346 | 3.66 |
| Portuguese | 333 | 3.52 |
| Slovakia | 414 | 4.38 |
| Slovenia | 329 | 3.48 |
| Chile | 343 | 3.63 |
| Türkiye | 461 | 4.87 |
| New Zealand | 459 | 4.85 |

As shown in Table 1., among the 24 OECD member countries in the study, Canada has the highest number of participants, representing 9.44% (893 students). In contrast, Sweden has the lowest participation rate at 3% (284 students). The participation rates for the other countries range from 5.69% to 3.01%, corresponding to 538 and 285 students, respectively.

Organization of Data

The assessment framework of the TIMSS 2015 application consists of achievement tests in mathematics and science and questionnaires that collect information about the educational and social environments that affect student achievement. Achievement tests aim to measure students' knowledge and skills in mathematics and science. Multiple-choice and open-ended questions are used in TIMSS. Multiple-choice questions offer four options. Only one of these options is correct. The number of wrong answers does not affect the correct answers. In open-ended questions, students form their answers by making explanations, making inferences based on some data, or drawing shapes. In this question type, scoring is done according to the scoring key specially determined for each question. In TIMSS 2015, students at the 4th-grade level participate in achievement tests, which are given 36 minutes for each section, and then complete questionnaires, which are given 30 minutes (TIMSS, 2015).

In this study, out of 14 test booklets used in TIMSS 2015, booklet number 7, which was administered to all countries, was selected for analysis. A total of 25 questions, 10 of which were open-ended and 15 of which were multiple-choice items, were included in the analysis. For multiple-choice questions, each correct answer was recoded as '1'



and each incorrect answer as '0'. In the open-ended questions, the answers identified as 'correct answers' were recoded as '1'. Answers labeled 'wrong,' 'partially correct,' 'inaccessible,' or 'omitted/invalid' were recoded as '0'.

Data Analysis

Before data analysis, the data set was prepared by re-coding items to make it suitable for analysis. First, measures of central tendency were calculated to test basic assumptions. The KR-20 reliability coefficient was then computed for internal consistency, followed by the assessment of skewness and kurtosis coefficients to examine the normality of the data set. The results, including measures of central tendency (mean, mode, median), standard deviation, range, kurtosis, skewness coefficients, and the KR-20 internal consistency coefficient, are presented in Table 2.

Table 2.

Test Statistics, Normality Tests, and Reliability Coefficients for Countries' Mathematics Test Scores

| Countries | n | Mean | Mode | Median | df | Range | Skewness | Kurtosis | KR-20 |
|----------------|-----|-------|------|--------|------|-------|----------|----------|-------|
| Germany | 285 | 11.86 | 13 | 212 | 4.56 | 23 | .155 | -.620 | .79 |
| USA | 720 | 14.24 | 17 | 14 | 5.24 | 24 | -.085 | -.701 | .84 |
| Australia | 431 | 12.31 | 15 | 12 | 5.24 | 24 | .019 | -.733 | .84 |
| Belgium | 384 | 14.50 | 16 | 15 | 3.93 | 21 | .013 | -.194 | .70 |
| Czech Republic | 376 | 12.54 | 11 | 13 | 4.86 | 24 | -.024 | -.512 | .81 |
| Finland | 357 | 13.81 | 17 | 14 | 4.45 | 24 | -.358 | -.244 | .78 |
| Holland | 357 | 13.81 | 17 | 14 | 4.45 | 24 | -.358 | -.244 | .78 |
| England | 321 | 13.56 | 14 | 14 | 3.96 | 20 | -.189 | -.324 | .70 |
| Ireland | 310 | 14.32 | 19 | 15 | 4.79 | 23 | -.243 | -.584 | .80 |
| Spain | 538 | 12.17 | 15 | 12 | 4.44 | 23 | -.060 | -.631 | .77 |
| Sweden | 284 | 12.36 | 13 | 12 | 5.06 | 23 | .110 | -.800 | .83 |
| Italy | 307 | 10.93 | 10 | 11 | 4.32 | 23 | .136 | -.243 | .75 |
| Japan | 314 | 18.24 | 20 | 19 | 4.63 | 22 | -.691 | -.004 | .82 |
| Canada | 893 | 12.01 | 11 | 12 | 4.78 | 24 | .181 | -.513 | .79 |
| Korea | 331 | 19.13 | 20 | 20 | 3.98 | 19 | -.924 | .610 | .80 |
| Hungary | 354 | 14.01 | 19 | 14 | 5.85 | 24 | -.031 | -1.055 | .87 |
| Norway | 286 | 11.15 | 8 | 11 | 4.47 | 23 | .162 | -.196 | .77 |
| Polonia | 346 | 13.81 | 15 | 14 | 4.93 | 25 | -.091 | -.548 | .82 |
| Portuguese | 333 | 13.57 | 11 | 14 | 4.86 | 21 | -.112 | -.706 | .81 |
| Slovakia | 414 | 10.66 | 8 | 10 | 4.95 | 21 | .180 | -.538 | .82 |

| | | | | | | | | | |
|-------------|-----|-------|----|----|------|----|-------|-------|-----|
| Slovenia | 329 | 12.48 | 9 | 12 | 4.87 | 24 | .069 | -.635 | .81 |
| Chile | 343 | 10.15 | 13 | 10 | 4.76 | 24 | .442 | -.113 | .82 |
| Türkiye | 461 | 11.38 | 12 | 11 | 5.36 | 24 | .159 | -.694 | .85 |
| New Zealand | 459 | 12.18 | 12 | 12 | 5.29 | 24 | -.017 | -.787 | .84 |

Table 2. shows that the measures of central tendency (mean, mode, median) for the mathematics achievement test scores are quite close across countries. Analysis of the normality values revealed that skewness coefficients were between -1 and +1 for all countries, while only Hungary had a kurtosis coefficient outside this range. Skewness and kurtosis coefficients within -1 and +1 indicate a normal distribution of test scores (Mertler & Vannatta, 2005). Thus, most of the data from the countries exhibit a distribution close to normal. In examining the KR-20 reliability coefficients, it was found that the scores ranged from .70 to .87, demonstrating acceptable reliability, as coefficients are expected to be between .70 and .80 (Nunnally & Bernstein, 1994).

Confirmatory factor analysis was conducted to determine whether the one-factor model for each country was confirmed, and the fit indices were examined. Afterward, the equality of covariance matrices was tested, followed by multi-group confirmatory factor analysis. The results of these analyses are presented in the findings section.

The analyses employed the maximum likelihood estimation method, which seeks to find the model that provides the highest probability of estimating the parameters. This is the most frequently applied statistical inference method for this type of data (White, 1982).

Before proceeding with the multi-group confirmatory factor analysis, the multicollinearity assumption was examined. Multicollinearity occurs when one independent variable is highly correlated with another to the extent that it may serve as a substitute. A multicollinearity problem is indicated when inter-item correlation coefficients exceed .90 (Çokluk, Şekercioğlu, & Büyüköztürk, 2018). In this context, inter-item correlation coefficients were calculated separately for each country, revealing that the correlation coefficients for the mathematics items ranged from .004 to .435. Thus, we conclude that there is no multicollinearity problem between the independent variables.

The study used χ^2 , χ^2/df , CFI, and SRMR values to compare models in the assessment of measurement invariance. The researchers also calculated the asymptotic covariance matrix, as the data set may deviate from normality in analyses conducted on large

samples, especially since the data are based on categorical scoring. The Satorra-Bentler χ^2 ($SB\chi^2$) value is necessary to calculate the T_s value, which determines whether there is a significant difference between the χ^2 values of the models (Satorra & Bentler, 2001). With the asymptotic covariance matrix included in the analysis, the $SB\chi^2$ value was used to evaluate model fit and to compare the models.

In terms of the hierarchical structure of the models, comparisons were made between nested models. Specifically, comparisons were conducted between the models: the structural invariance model (Model 1) and the weak factorial invariance model (Model 2), the weak factorial invariance model (Model 2) and the strong factorial invariance model (Model 3), and the strong factorial invariance model (Model 3) and the strict factorial invariance model (Model 4). Firstly, the difference in $\Delta\chi^2$ values and degrees of freedom (Δdf) between the two models was analyzed. The statistical significance of $\Delta\chi^2$ was assessed by referencing the critical value corresponding to the difference in degrees of freedom at the $p < .05$ significance level in the chi-square distribution table. After calculating the $SB\chi^2$ value, T_s values were determined to evaluate the significance of the differences. If the calculated T_s value exceeds the critical chi-square value, it indicates that the difference between the compared models is significant, suggesting that measurement invariance is not achieved. Conversely, if the T_s value is smaller than the critical chi-square value, it indicates that the difference between the models is not significant. Additionally, ΔCFI and $\Delta SRMR$ values were also considered in comparison.

In making model comparisons, specific cut-off values based on sample size were used as references. For the weak factorial invariance test between groups when the sample size is small ($n < 300$), the cut-off values are $\Delta CFI \leq .005$ and $\Delta SRMR \geq .025$. For the comparison of strong and strict factorial invariance in this sample size category, the cut-off values are $\Delta CFI \geq .005$ and $\Delta SRMR \geq .005$. When the sample size is sufficient ($n > 300$), the cut-off values for the weak factorial invariance test adjust to $\Delta CFI \geq .010$ and $\Delta SRMR \geq .030$, and for the strong and strict factorial invariance comparison, they are $\Delta CFI \geq .010$ and $\Delta SRMR \geq .010$ (Chen, 2007). These cut-off values help evaluate model fit across different sample sizes.

FINDINGS

This section presents the findings from the confirmatory factor analysis, the test of equality of covariance matrices, and the multi-group confirmatory factor analysis. The results from the confirmatory factor analysis are shown in Table 3.

Table 3.

Confirmatory Factor Analysis Results for Countries' Mathematics Achievement Test Scores

| Countries | SB χ^2 (df) | χ^2 /df | CFI | NNFI | SRMR | RMSEA |
|----------------|------------------|--------------|------|------|------|-------|
| Germany | 266.95(275) | 0.97 | 1.00 | 1.00 | .046 | .000 |
| USA | 442.62(275) | 1.61 | .98 | .98 | .036 | .029 |
| Australia | 321.84(275) | 1.17 | .99 | .99 | .039 | .020 |
| Belgium | 281.24(275) | 1.02 | .99 | .99 | .043 | .008 |
| Czech Republic | 369.67(275) | 1.34 | .96 | .96 | .046 | .030 |
| Finland | 284.54(275) | 1.03 | 1.00 | .99 | .042 | .010 |
| Holland | 384.82(275) | 1.40 | .89 | .88 | .055 | .035 |
| England | 379.10(275) | 1.38 | .96 | .96 | .053 | .037 |
| Ireland | 283.34(275) | 1.03 | 1.00 | 1.00 | .045 | .010 |
| Spain | 345.81(275) | 1.26 | .97 | .97 | .039 | .022 |
| Sweden | 334.63(275) | 1.22 | .97 | .97 | .049 | .028 |
| Italy | 343.92(275) | 1.25 | .95 | .94 | .052 | .029 |
| Japan | 304.76(275) | 1.11 | .99 | .99 | .050 | .019 |
| Canada | 478.40(275) | 1.74 | .96 | .96 | .035 | .029 |
| Korea | 314.86(275) | 1.14 | .98 | .98 | .050 | .021 |
| Hungary | 331.81(275) | 1.21 | .99 | .99 | .042 | .024 |
| Norway | 291.54(275) | 1.06 | .99 | .99 | .048 | .015 |
| Polonia | 381.13(275) | 1.39 | .96 | .96 | .048 | .033 |
| Portuguese | 390.05(275) | 1.42 | .95 | .94 | .050 | .035 |
| Slovakia | 442.60(275) | 1.61 | .95 | .95 | .048 | .038 |
| Slovenia | 328.14(275) | 1.19 | .98 | .98 | .047 | .024 |
| Chile | 312.96(275) | 1.14 | .99 | .98 | .044 | .020 |
| Türkiye | 330.73(275) | 1.20 | .99 | .99 | .038 | .021 |
| New Zeland | 448.28(275) | 1.63 | .97 | .96 | .045 | .037 |

Table 3. shows that the χ^2 /df ratio of the mathematics achievement test scores for all participating countries is below 3. The analysis indicates that the CFI values for all countries, except the Netherlands, are above .90; the CFI value for the Netherlands is .89. Similarly, the NNFI value for all countries included in the analysis, except the Netherlands, is .90 or above, while the NNFI value for the Netherlands is .88. In terms of SRMR values, 21 countries have values below .05, while the SRMR values for the Netherlands, England, and Italy are .55, .53, and .52, respectively, which are significantly above .05. Since RMSEA values are below .05 for all countries, we can conclude that, generally, each model for the 24 countries is confirmed according to the results of the confirmatory factor analysis.

The equality of the covariance matrices of the mathematics achievement test scores of the OECD member countries was tested before proceeding to the multi-group confirmatory factor analysis. As a result of the analysis, $SB\chi^2(7475)=13170.6$, $p=.000$,

$\chi^2/df=1.76$, $RMSEA=.044$, $GFI=.90$, $CFI=.92$ and $SRMR=.081$. According to this, the bd ratio is below 3.

According to the analysis, $SB\chi^2/df$ ratio is below 3, the $RMSEA$ value is below .05, the GFI value is equal to .90, the CFI value exceeds .90, and the $SRMR$ value is above .05. Overall, this indicates that the covariance matrices between countries are consistent.

Finally, Table 4. presents the results of the multi-group confirmatory factor analysis conducted to determine whether measurement invariance of the TIMSS 2015 4th-grade mathematics test scores was ensured across OECD member countries.

Table 4.

Findings of the Multi-Group Confirmatory Factor Analysis for TIMSS 2015 4th-Grade Mathematics Achievement Test Scores of OECD Member Countries

| | $SB\chi^2(df)^1$ | $\Delta\chi^2(\Delta df)$ | χ^2/df | $\Delta\chi^2/(\Delta df)$ | CFI | ΔCFI | SRMR | $\Delta SRMR$ |
|---------------------|------------------|---------------------------|-------------|----------------------------|-----|--------------|------|---------------|
| Model1 ^a | 15034.70(7750) | - | 1.9399 | - | .90 | - | .081 | - |
| Model2 ^b | 12495.26(7175) | 2539.44(575) | 1.7415 | .1984 | .93 | -.03 | .051 | .030 |
| Model3 ^c | 8381.66(6600) | 4113.6(575) | 1.2699 | .4716 | .98 | -.05 | .045 | .006 |
| Model4 ^d | 10845.62(7175) | -2463.96(-575) | 1.5116 | -.2417 | .95 | .03 | .079 | -.034 |

¹ $p<.05$, ^aStructural Invariance (Factor loadings, factor correlations, and error variances are constant), ^bWeak Factorial Invariance (Factor loadings are free, factor correlations and error variances are constant), ^cStrong Factorial Invariance (Factor loadings and error variances are free, factor correlations are constant), ^dStrict Factorial Invariance (Error variances are free, factor loadings and factor correlations are constant)

Firstly, the structural invariance model was tested to examine whether the factor structures of the countries to be compared are similar. According to the analysis, the $SB\chi^2$ and degrees of freedom ratio (χ^2/df) were below 2, the CFI value was .90, and the $SRMR$ value was slightly above .08. This indicates that the fit indices for Model 1 are generally acceptable, validating the model. Thus, it can be concluded that the structural invariance model is achieved.

When comparing the structural invariance (Model 1) and weak factorial invariance (Model 2) models, it was determined that the $\Delta\chi^2$ and Δdf ratios improved. The CFI value differed significantly ($\Delta CFI<-.01$), and there was a significant change in the $SRMR$ value ($\Delta SRMR>.025$). To assess the significance of the $\Delta\chi^2$ and Δdf improvements, the T_s value was calculated at 2972.648, which exceeded the critical value in the χ^2 distribution table, $\chi^2_{diff}(575)=631.893$, $p<.05$. Overall, there is a significant difference between the structural invariance and weak factorial invariance models, as indicated by significant differences in all fit indices.

In the comparison of the weak factorial invariance (Model 2) and strong factorial invariance (Model 3) models, the $\Delta\chi^2$ and Δdf ratios also improved. The T_s value was

calculated at 7194.932 and was above the critical value in the χ^2 distribution table, $\chi^2_{diff}(575)=631.893$, $p<.05$). This indicates a significant difference between the weak and strong factorial invariance models. Additionally, for Models 2 and 3, the CFI value showed a significant difference ($\Delta CFI<-.01$), but there was no significant difference in the SRMR value ($\Delta SRMR<.01$).

Finally, when comparing the strong factorial invariance (Model 3) and strict factorial invariance (Model 4) models, the $\Delta\chi^2$ and Δdf ratios worsened. However, the changes in CFI ($\Delta CFI>-.01$) and SRMR ($\Delta SRMR>.01$) values were not significant. Thus, it can be concluded that the difference between the strong factorial invariance model and the strict factorial invariance model is not significant, as there is no significant difference in two of the three fit indices.

Based on these findings, the strong factorial invariance model is the best-fitting model among the four. Therefore, it suggests that the structure of the mathematics achievement test is not equal across OECD countries; in other words, measurement invariance is not achieved.

DISCUSSION AND CONCLUSION

This study examined whether the scores obtained from the TIMSS 2015 mathematics achievement test provide measurement invariance for OECD member countries. Firstly, we examined the basic assumptions that the data set should meet, calculating measures of central tendency, standard deviation, range, and reliability coefficients. The findings indicate that the distribution was close to normal, with reliability coefficients exceeding .70 for each country, suggesting that the internal consistency of the test items was at an acceptable level.

We performed confirmatory factor analysis to determine whether the measurement models for the country groups were confirmed. The analysis results showed that the goodness of fit indices were sufficient, confirming the models for the countries. Before proceeding to measurement invariance analyses, we tested the equality of the covariance matrices for the country groups. The test results indicated a good fit between the covariance matrices of the countries.

Multi-group confirmatory factor analysis was used to test the measurement invariance of scores obtained from the fourth-grade mathematics achievement test in TIMSS 2015 by country. According to the findings, the fit indices for Model 1 were generally acceptable, confirming that the structural invariance model was met. The analysis involved comparisons between nested models; first, the structural invariance model



(Model 1) was compared with the weak factorial invariance model (Model 2). Then, Model 2 was compared with the strong factorial invariance model (Model 3). The fit indices improved significantly in both comparisons. However, the comparison between the strong factorial invariance model (Model 3) and the strict factorial invariance model (Model 4) showed that the fit deteriorated compared to Model 3. Therefore, it is concluded that the strong factorial invariance model (Model 3) is the best-fitting model but does not provide measurement invariance for the mathematics achievement test scores of OECD member countries.

According to the findings obtained for the country groups, this study is similar to Ayvalli's (2016) investigation of the measurement invariance of the mathematics literacy test from the PISA 2012 application among OECD member countries. In Ayvalli's study, measurement invariance across countries was achieved at the level of strong factorial invariance. Similarly, the findings were consistent with Kıbrıslıoğlu's (2015) study, which examined the measurement invariance of the PISA 2012 mathematics learning model across countries. However, in Kıbrıslıoğlu's study, measurement invariance was not achieved; the measurement model was confirmed only at the structural invariance level. Additionally, Karakoç Alatl (2016) found that measurement invariance was not achieved in the scores obtained from the PISA 2012 mathematics and science literacy tests and reading skills tests among groups from Australia, Shanghai-China, Turkey, and France, particularly concerning the language variable.

Similar results were observed in studies involving the Turkish sample. These studies investigated the invariance of scores collected from scales used in PISA and TIMSS applications based on statistical region, gender, or year variables, and found that measurement invariance was achieved in most cases (Uzun & Öğretmen, 2010; Uyar & Doğan, 2014; Başusta & Gelbal, 2015; Ayvalli, 2016; Kıbrıslıoğlu Uysal & Akın Arıkan, 2018). In contrast, similar results in cross-country studies indicated that evidence of measurement invariance from the scales used in PISA and TIMSS applications among different country groups was largely not obtained (Kıbrıslıoğlu, 2015; Ayvalli, 2016; Karakoç Alatl, 2016; Şekercioğlu & Koğar, 2018). This raises questions about the assumption that tests administered at the international level are perceived equally by participants from different countries and that the measurement process is consistent. The confirmation of this assumption in most studies involving the Turkish sample suggests that culture is a crucial factor influencing this

discrepancy. Therefore, it is essential to be meticulous when adapting assessment studies conducted across multiple countries to different languages and cultures.

When considering the results in general, the inability to ensure measurement invariance of the fourth-grade mathematics achievement test from the TIMSS 2015 application across countries has called into question the rankings of the participating countries and the interpretations made based on those rankings.

Declaration of Contribution of Researchers

This study is based on the first author's master's thesis, which was completed under the supervision of the second author.

Conflict Declaration

There is no conflict between the authors.



REFERENCES

- Ayvallı, M. (2016). PISA 2012 matematik okuryazarlığı testinin ölçme değişmezliğinin incelenmesi. (Unpublished master's thesis). Akdeniz University.
- Başusta, N.B., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30(4), 80-90.
- Bayır, E., Çakıcı, Y., & Atalay, Ö. (2016). Fen bilimleri öğretmenlerinin bilimin doğasına ilişkin görüşleri: Bilişsel harita örneği. Kastamonu Eğitim Dergisi, 24(3), 1419-1436.
- Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement equivalence. Psychological Bulletin, 105, 456-466.
- Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105(3), 456-466. doi:10.1037/0033-2909.105.3.456
- Çepni, S., Ayas, A., Johnson, D., & Turgut, F. (1997). Fizik öğretimi. Ankara: YÖK/Dünya Bankası Milli Eğitimi Geliştirme Projesi Hizmet Öncesi Öğretmen Eğitimi.
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural equation modeling, 14(3), 464-504. doi:10.1080/10705510701301834
- Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233-255. Doi:10.1207/S15328007SEM0902_5
- Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2018). Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları. Pegem Akademi.
- Ergül, H. (1999). Uzaktan öğretimde kalite verimlilik ve üretkenlik. Anadolu Üniversitesi Açık Öğretim Fakültesi Kurgu Dergisi, 16, 283-296.
- Gierl, M.J. (2000). Construct equivalence on translated achievement tests. Canadian Journal of Education, 25(4), 280-296. doi:10.2307/1585851
- Horn, J.L. & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18(3), 117-144. doi:10.1080/03610739208253916
- International Association for Evaluation of Educational Assessment (2015). TIMSS 2015: The trends in international mathematics and science study. Retrieved from https://www.iea.nl/fileadmin/user_upload/Studies/TIMSS_2015/TIMSS_2015.pdf on July 25, 2015.
- Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426. doi:10.1007/BF02291366
- Karasar, N. (2005). Bilimsel araştırma yöntemi, Nobel Yayıncılık.

- Karakoç Alatl, B. (2016). Uluslararası öğrenci değerlendirme programı (PISA-2012) okuryazarlık testlerinin ölçme değişmezliğinin incelenmesi. (Unpublished master's thesis). Ankara University.
- Kıbrıslıoğlu, N. (2015). PISA 2012 matematik öğrenme modelinin kültürlere ve cinsiyete göre ölçme değişmezliğinin incelenmesi: Türkiye-Çin (Şangay)-Endonezya örneği. (Unpublished master's thesis). Hacettepe University.
- Kıbrıslıoğlu Uysal, N. & Akın Arıkan, Ç. (2018). Measurement invariance of science self-efficacy scale in PISA. *International Journal of Assessment Tools in Education*, 5(2), 325-338.
- MEB (2010). Uluslararası öğrenci değerlendirme programı (PISA) 2009 ulusal ön raporu. Millî Eğitim Bakanlığı Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı, Ankara. Retrieved from <https://pisa.meb.gov.tr/www/raporlar/icerik/5> on July 25, 2015.
- MEB (2014). TIMSS 2011 ulusal matematik ve fen raporu: 8. sınıflar. Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü. Retrieved from https://timss.meb.gov.tr/meb_iys_dosyalar/2022_03/07135958_TIMSS-2011-8-Sinif.pdf on July 25, 2015.
- Mark, B.A. & Wan, T.T.H. (2005). Testing measurement equivalence in a patient satisfaction instrument. *Western Journal of Nursing Research*, 27(6), 772-787.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. doi:10.1007/BF02294825
- Mertler, C.A. & Vannatta, R.A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation (Third Edition)*. Pyrczak.
- Milfont, T.L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111-121. doi:10.21500/20112084.857
- Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory. (Third Edition)*. New York: McGraw-Hill, Inc.
- Putnick, D.L. & Bornstein, M.H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. doi: 10.1016/j.dr.2016.06.004
- Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press, Boca Raton.
- Satorra, A. & Bentler, P.M. (2001). A scaled difference chi-square test statistics for moment structure analysis. *Psychometrika*, 66(4), 507-514. doi:10.1007/BF02296192
- Schoot, R., Lugtig, P. & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486-492. doi:10.1080/17405629.2012.686740
- Şekercioğlu, G. (2018). Measurement invariance: Concept and implementation. *International Online Journal of Education and Teaching (IOJET)*, 5(3), 609-634.
- Şekercioğlu, G. ve Koğar, H. (2018). The examination of measurement invariance and differential item functioning of PISA 2015 cognitive tests in terms of the commonly used languages. *Novitas-Royal (Research on Youth and Language)*. 12(2), 152- 172.



- Uluslararası Matematik ve Fen Eğilimleri Araştırması TIMSS (2015). TIMSS 2015 ulusal matematik ve fen ön raporu. Retrieved from http://timss.meb.gov.tr/wpcontent/uploads/TIMSS_2015_Ulusal_Rapor.pdf September 13, 2018.
- Uyar, Ş. & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi*, 2, 30-43.
- Uzun, B. & Öğretmen, T. (2010). Fen başarısı ile ilgili bazı değişkenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi. *Eğitim ve Bilim*, 35(155), 26-35.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometria*, 50(1), 1-25.
- Wu, A.D., Li, Z. & Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multigroup confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, 12(3), 1-26.
- Vandenberg, R.J. & Lance, C.E. (2000). A review and synthesis of the MI literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70. doi:10.1177/109442810031002.