Bartın University Journal of Faculty of Education BUEFAD
2023, Volume 12, Issue 3, 579-593
dergipark.org.tr/buefad
DOI: 10.14686/buefad.1104446

An Examination of TIMSS 2015 Science Affective Factors with Regard to Gender and Regions

Mehmet Can DEMİR [a*], ***Selahattin GELBAL*** [b]

a Res. Asst., Bartın University, https://orcid.org/0000-0001-7849-7078 * mehmetdemir@bartin.edu.tr
b Prof. Dr., Hacettepe University, https://orcid.org/0000-0001-5181-7262

Research Article
Received: 16.4.2022
Revised: 28.11.2022
Accepted: 23.1.2023

Abstract

Various international and national tests are applied to assess curricula across the world. About 250.000 students' knowledge and skills are tested via Trends in International Mathematics and Science Study (TIMSS). Many international and national comparisons are made with the test results. However, to compare the test results of such a big crowd of people, the test results must not differ because of any other variable but achievement; in short, they must represent the same results of measurement. In the presence of measurement invariance, the structure in question will be similar in all subgroups. In light of this information, the purpose of the study is to examine the measurement invariance of the model of the science affective factors of the eighth graders who participated in TIMSS 2015 Turkey, which is constructed via structural equation modeling, among Nomenclature of Territorial Units for Statistics – Level 1 regions and genders in Turkey, through multi-group confirmatory factor analysis. The sample of this research consists of 5344 students. As a result of this study, the model met strict invariance for genders and scalar invariance for regions.

Keywords: TIMSS, science affective factors, structural equation modeling, measurement invariance, multi-group confirmatory factor analysis

TIMSS 2015 Fen Duyuşsal Özelliklerinin Cinsiyete ve Bölgelere Göre İncelenmesi

Öz

Dünya çapında, eğitim programlarının değerlendirilmesi amacıyla birçok uluslararası ve ulusal sınav uygulanmaktadır. Bu sınavlardan biri olan Uluslararası Matematik ve Fen Eğilimleri Araştırması ile yaklaşık 250.000 öğrencinin bilgi ve becerileri test edilmektedir. Test sonuçları ile uluslararası ve ulusal düzeyde birçok karşılaştırma yapılmaktadır. Fakat bu kadar büyük bir kitlenin sonuçlarının anlamlı bir şekilde karşılaştırılabilmesi için sınav sonuçlarının başarıdan farklı bir değişkenden ötürü farklılık göstermemesi, aynı ölçüm sonuçlarını yansıtması gereklidir. Ölçme değişmezliğinin varlığında söz konusu yapının tüm alt gruplarda benzerdir. Bundan hareketle, araştırma kapsamında TIMSS 2015 Türkiye uygulamasına katılan 8. sınıf öğrencilerinin fen duyuşsal özellik değişkenlerinden öğrenci anketine verilen yanıtlar kullanılarak Yapısal Eşitlik Modellemesi ile bir model oluşturulmuştur. Daha sonra bu modelin cinsiyet ve İstatistiki Bölge Birimleri Sınıflaması Düzey 1'e göre Türkiye'de yer alan bölgelere göre ölçme değişmezliği, çok gruplu doğrulayıcı faktör analizi ile incelenmiştir. Araştırmanın örneklemi TIMSS 2015 Türkiye uygulamasına katılan 5344 öğrenciden oluşmaktadır. Çalışma sonucunda, oluşturulan modelin cinsiyetler göre katı değişmezlik, bölgeler arasında ölçek değişmezliği koşulunu sağladığı gözlenmiştir.

Anahtar kelimeler: TIMSS, fen duyuşsal özellikleri, yapısal eşitlik modellemesi, ölçme değişmezliği, çok gruplu doğrulayıcı faktör analizi

To cite this article in APA Style:

Demir, M. C. & Gelbal, S. (2023). An examination of TIMSS 2015 science affective factors with regard to gender and regions. *Bartın University Journal of Faculty of Education, 12*(3), 579-593. https://doi.org/10.14686/buefad.1104446

INTRODUCTION

Educational achievement is critical in determining the position in international competition and social development. With the facilitating effect of technological developments, many studies (e.g., ABIDE, ÖBBS, TIMSS, PISA, PIRLS, etc.), including national and international comparisons, are carried out in education today. As a result of these studies, countries have information about the level of educational achievement. Governments also have access to information about the status of different groups in terms of various characteristics (e.g., socioeconomic status, culture, gender, etc.) within themselves. With this information, they can improve/develop their education systems from different perspectives.

The Trends in International Mathematics and Science Study (TIMSS), which is organized by The International Association for the Evaluation of Educational Achievement (IEA), is an international survey that evaluates students' achievements in science and mathematics (Mullis & Martin, 2013). With TIMSS, at the international level, the knowledge and skills of fourth and eighth-grade students acquired in mathematics and science courses are evaluated, and it is aimed to determine the extent to which students can use what they learned at school in their daily life and assess their state of having high-level cognitive skills. In addition to mathematics and science achievement tests, student, teacher, school administrator, and parent questionnaires are also included in the study. Much detailed information about students' socioeconomic levels, educational opportunities at home and school, and school climate is obtained from these student questionnaires. The student questionnaire used in TIMSS contains items related to affective characteristics that affect achievement, such as interest, motivation, attitude, school belonging, peer bullying, and the value given to the course (Mullis et al., 2016).

In studies in social and educational sciences, comparisons of the construct (e.g., achievement) between different groups in the sample (e.g., gender, parental education level, socioeconomic status) are frequently included (Cronbach & Meehl, 1955). With the data obtained from TIMSS 2015, many comparisons and evaluations were made at the national and international levels regarding different variables. When the results of the TIMSS 2015 Turkey application are analyzed, it is seen that there are significant differences in terms of scores between 12 regions in the Nomenclature of Territorial Units for Statistics (NUTS) Level 1. Likewise, when compared by gender, it was seen that there were differences between the scores of male and female students. Since individuals have different characteristics, it is a natural result. However, it is not always true that the only reason for the difference between the scores is that the individuals are different, which is explained only by this (Başusta, 2010). Because the reason for the difference may be the construct itself or the measurement tool itself. If a measurement tool is developed to be applied to different groups, it is necessary to prove that the psychometric properties of the observed indicators are equivalent in all groups. It is possible for the comparisons made to be correct if individuals with the same ability level in different subgroups in the sample get equal scores regarding the measured variable (Schmitt & Kuljanin, 2008). When this equivalence is achieved, it is concluded that the scores obtained from the scales are comparable between the groups; that is, measurement invariance is ensured. Measurement invariance is that the latent variables have the same relationship with the observed variables across all groups (Widaman & Reise, 1997). Measurement invariance refers to whether an instrument (e.g., a scale) can measure a latent trait similarly across different population subgroups or over multiple points in time. Measurement invariance is not related to the characteristics of the individuals from whom the measurement is obtained but to the tool itself. Comparisons between groups for which the level of measurement invariance has not been proven may not always yield accurate results. Therefore, when measurement invariance cannot be confirmed, it is not possible to make meaningful comparisons and evaluations between groups (Öğretmen, 2006).

Measurement invariance is a concept related to construct validity, which is one of the methods of collecting proof of validity. Construct validity is about determining the boundaries of the structure that is accepted to exist and trying to be measured and making it observable. In the presence of measurement invariance, it is assumed that the structure in question will be similar in all subgroups. If the measurement invariance is not ensured, the validity of the results will not be proven (Gregorich, 2006). In this case, the accuracy of all comparisons and interpretations to be made with the results will be doubtful. Therefore, measurement invariance studies are of critical importance as they provide substantial evidence about the validity of the results.

In measurement invariance studies, a measurement model is established between observed and latent variables; then, it is examined whether this model has the same structure in different groups that are compared between the model. The fact that the model has the same structure in different groups means that the factor loadings, the correlations between the factors, and the error variances of the factors are equal (Jöreskog & Sörbom, 1993). When the studies conducted to date for measurement invariance are examined, two different approaches

are seen: Item Response Theory (IRT)-based and Structural Equation Modeling (SEM)-based approaches (Reise et al., 1993; Raju et al., 2002). A theoretical measurement model is established in both approaches, but due to the methodological differences in modeling, the approaches have advantages and disadvantages (Meade & Lautenschlager, 2004).

In IRT-based measurement invariance studies, a log-linear model is created, and the intergroup comparison of the model is multidimensional. Invariance is examined through the model's functional differences at the item and test levels. In cases where invariance is not ensured, the model parameters are estimated and scaled for each of the subgroups, provided that the fit between the model and the data is ensured. Then, the item/items that are the source of invariance are determined by comparing the item characteristic curves (Karakoç Alatlı, 2016). In SEM-based studies, multi-group confirmatory factor analysis (MGCFA), invariance of mean and covariance structures (MACS), and equality of covariance structures are considered (Yandı et al., 2017). When the studies are examined, it is seen that the MGCFA method is used more frequently in measurement invariance studies.

Measurement invariance test through MGCFA is usually conducted by testing four hierarchical models sequentially. The four steps of measurement invariance, or in other words, the models tested, are configural invariance, metric invariance, scalar invariance, and strict invariance (Bialosiewicz et al., 2013). These steps are summarized below.

- The configural invariance step tests whether the specified model is the same in the groups (Kline, 2011). However, if this equality is not achieved, it means that the factor structure is not the same across the groups; that is, the items do not measure the same structure in different groups, so there is no need to make comparisons between groups and move on to the next step of invariance (Vandenberg & Lance, 2000).

- In the metric invariance step, in addition to the equality of the structures in the groups, it is tested whether the item factor loadings are equal (Millsap & Olivera-Aguilar, 2012). Latent variables have an effect on the observed variables with the relation of factor loadings (Bollen, 1989). Therefore, the structure cannot be measured invariably in groups where the factor loadings indicating the degree of influence of the latent variable on the observed variable are not the same. The unequal group factor loadings indicate that the individuals do not interpret the items similarly (Bialosiewicz et al., 2013). Since the factor loadings between the groups are equal, the factor variance and covariance of the groups are suitable for comparison. However, the source of the difference in the means of the groups is still unclear. (Millsap & Olivera-Aguilar, 2012).

- In the scalar invariance step, in addition to the equality of factor loadings, it is tested whether the intercepts of the items are equal across the groups (Millsap & Olivera-Aguliar, 2012). Ensuring this equality means that the means of the factors and observed variables are comparable (Gregorich, 2006). In other words, the items explain the factors at the same level in different groups. Likewise, since scalar invariance is ensured, it can be interpreted that the source of the difference between the means of the groups is the latent variable itself, not another variable other than the latent variable (Başusta & Gelbal, 2015).

- In the strict invariance step, in addition to the equality of the item intercepts, it is tested whether the error variances of the items are equal across the groups (Widaman & Reise, 1997). Ensuring strict invariance means that the variances and covariances of the observed variables are comparable between groups (Gregorich, 2006). However, since this model is based on the condition of equality of item error variances, it takes work to meet invariance in practice because the increase in the variance of the latent variable causes an increase in the item error variances (Widaman & Reise, 1997).

The factors affecting student achievement have been investigated in many different national and international studies carried out to date. Then, comparisons of student achievement in terms of these factors (e.g., socioeconomic status, student characteristics, teacher characteristics, etc.) are included. Although the primary purpose of TIMSS is to improve science and mathematics teaching, it is seen that comparisons are made in terms of student achievement and affective characteristics in the reports related to the research (Mullis et al., 2016). However, making comparisons and accepting the results as they are without examining whether the comparisons are feasible can lead to erroneous decisions. Before making comparisons, the measurement invariance of the

constructs should be examined. Because of the measurement invariance study, precise information is obtained about the factors to be used for comparison and to what extent the comparisons can be made.

Evaluating students' course achievement independently of their affective characteristics is impossible. The emotional tendencies of the students toward the lessons constitute the affective characteristics of the students (Bloom, 2012). Affective traits include interest, attitude, anxiety, motivation, self-esteem, etc. can be interpreted as a combination of variables. Previous studies have demonstrated that the affective characteristics of a course affect the success of that course. According to Bloom (2012), approximately 25% of the differences in student achievement are due to affective characteristics. Due to the significant effect of affective factors on the difference in course achievement, it is necessary to consider the effect of differentiation in terms of affective attributes in cases where achievement comparisons are made.

When the reported TIMSS 2015 Turkey National Preliminary Report is analyzed, there is a 70-point difference between the West Marmara Region, which has the highest average in eighth-grade science scores, and the Middle East Anatolia Region, which has the lowest average (MoNE, 2016). When the scores of male and female students were examined, it was seen that the mean of male students was 484, the mean of female students was 503, and the difference was statistically significant. When the literature is examined, it is seen that measurement invariance studies for test and survey scores obtained in TIMSS and other national and international evaluation studies are conducted (Akyıldız, 2009; Bahadır, 2012; Ercikan & Koh, 2005; Gülleroğlu, 2017; Karakoç Alatlı, 2016; Kıbrıslıoğlu, 2015; Ölçüoğlu & Çetin, 2016; Uyar, 2011; Uzun, 2008; Wu et al., 2007; Yandı et al., 2017).

Based on the difference between gender and region average scores, in this study, the measurement invariance of the science affective trait model, established using the student questionnaire items in the TIMSS 2015 Turkey application, was tested between genders and regions. The study aimed to examine the measurement invariance of the eighth-grade science affective trait model between genders and regions and provide evidence about the validity of the scores.

Research Question

Does the science affective trait model specified using the student questionnaire according to TIMSS 2015 data meet measurement invariance between genders and NUTS Level 1 regions of Turkey?

METHOD

Research Design

The study is a descriptive study to prove evidence about the validity of the TIMSS 2015 cycle, to make evaluations, and to reveal possible relationships (Büyüköztürk et al., 2012). Moreover, quantitative research methods have been employed as answers to research problems are sought by using quantitative data.

Participants of the Study

The target audience of TIMSS is all students enrolled in formal education at the fourth and eighth-grade levels. Approximately 600,000 students from 60 countries participated in TIMSS 2015 (MoNE, 2016). In TIMSS, the school sample is determined by the stratified random sampling method. For the TIMSS 2015 application, schools were selected by stratified random sampling method using NUTS Level 1, school type, location of schools, and administrative forms of schools in the first stage. In the second stage, the students who will participate in the application in these schools were randomly determined.

The TIMSS 2015 Turkey application population consists of all students studying in Turkey at these two grade levels. However, since this research was limited to eighth-grade students, the population of this research consisted of 1,187,893 students (MoNE, 2016). Six thousand seventy-nine students representing eighth-grade students in 12 regions and rural areas by NUTS Level 1 participated in the TIMSS 2015 cycle. The sample consists of 2943 female (48.4%) and 3136 male (51.6%) students. Descriptive statistics of students by region are shown in Table 1.

Table 1. Number of Students by Regions

Regions	n	Percent
Istanbul (TR1)	913	17.08
West Marmara (TR2)	214	4.00
Aegean (TR3)	607	11.36
East Marmara (TR4)	577	10.80
West Anatolia (TR5)	341	6.38
Mediterranean (TR6)	570	10.67
Central Anatolia (TR7)	274	5.13
West Black Sea (TR8)	195	3.65
East Black Sea (TR9)	252	4.72
Northeast Anatolia (TRA)	208	3.89
Central East Anatolia (TRB)	327	6.12
Southeast Anatolia (TRC)	866	16.21
Total	5344	100

According to Table 1, the number of students participating in the application is the highest in Istanbul and the lowest in the Western Black Sea Region, in proportion to the population of the regions. Since the data on the rural area covers all rural areas in Turkey, the study was conducted with the data of the students in regions other than the rural areas. Accordingly, the analyses were carried out with the data of 5344 students in 12 regions.

Data Collection

The TIMSS 2015 cycle was mainly carried out with mathematics and science achievement tests and affective characteristics questionnaires for these courses. The items in the achievement tests were prepared for the achievements within the scope of the TIMSS Mathematics and Science Evaluation Framework. In the student questionnaires, there are items about the socioeconomic status of the family, the educational resources at home and at school, their attitudes towards mathematics and science lessons, and the school climate. Achievement tests were administered in two sessions of 45 minutes each, and questionnaires were administered in one session of 30 minutes. In addition to student tests and questionnaires, TIMSS also includes questionnaires prepared for the administrators of the schools participating in the application and the science and mathematics teachers of the classes. This study used the answers to the science student questionnaire in the TIMSS 2015 Turkey eighth-grade application as data. Only secondary data were used, and no other data was collected.

Data Analysis

Data analysis was completed in three steps. First, the suitability of the data for the analyses was tested. The data was split into two equal parts by random, to use in exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). This was for avoiding overfitting in EFA and CFA (Fokkema & Greiff, 2017). Then, a science affective trait model was specified using EFA, and the model was tested by CFA. Finally, the measurement invariance of the model according to gender and region was tested via MGCFA. $\Delta CFI \leq .01$ value between models (Cheung & Rensvold, 2002) was used as cutoff criteria in MGCFA. All the steps in the analysis process are explained in detail below.

Research Ethics

The approval of the Ethics Commission of Hacettepe University Senate was obtained for ethical compliance with the research procedures. All data except for the region codes are retrieved from the official website of the IEA (https://www.iea.nl/data-tools/repository/timss). Data of regional codes were obtained with permission from the Ministry of National Education, General Directorate of Assessment and Examination Services.

FINDINGS

Testing of Model Assumptions

The data from 208 students in the sample could not be reached for various reasons. The data from 29 students who did not answer a sufficient number of questions, including questions about demographic information

such as gender and age, were excluded from the study. The missing rate for the variables in the study is between 0.2% and 2.0%. Since the missing rate is less than 5% for each variable and all combinations of the variables, the missing data was completed using the Expectation Maximization (EM) algorithm method.

After dealing with the missing data, outliers were checked. As a result of univariate outlier analysis, 93 data whose Z scores were not in the range of ±4 were removed. As a result of the multivariate outlier analysis, 557 data with Mahalanobis distances higher than the critical $\chi 2$ value were removed. After the extreme values were cleared, the study continued with the data of 4457 students. The distribution of the number of students by gender and region in the final version of the dataset is similar, as the missing data is not cleaned.

In the analysis of univariate normality, it was observed that the Z scores of skewness and kurtosis coefficients of some items were not within the range of ±1.96. Then, Mardia's test was conducted to examine multivariate normality, and both test statistics of skewness and kurtosis were statistically significant ($p<.001$; $p<.001$). However, the fact that the skewness and kurtosis coefficients are significantly different from 0 does not make a difference in large samples in such a way as to impair normality significantly (Tabachnick & Fidell, 2013). Although the sample size was 4457, it is assumed that univariate and multivariate normality was not met.

When the correlation coefficients between the items were examined, no correlation higher than .90 was found. It was observed that tolerance values for all items were greater than .01, variance inflation factor (VIF) values were less than 10 and conditional index (CI) values were less than 30. These results show no multicollinearity and singularity problem in the data set. After this step, data was split into two equal parts randomly. However, 12 regions were combined into seven regions considering the characteristics of the regions since CFA estimation could not converge because very few people remained in the subgroups. The frequencies of the groups are given in Table 2.

Table 2. Number of Students by Regions After Split

Merged Regions	Regions	EFA Dataset		CFA Dataset	
		n	Percent	n (Merged)	Percent (Merged)
Marmara	Istanbul (TR1)	353	15.84	657	29.48
	West Marmara (TR2)	80	3.59		
	East Marmara (TR4)	238	10.68		
Aegean	Aegean (TR3)	270	12.12	295	13.23
Central Anatolia	West Anatolia (TR5)	139	6.24	281	12.61
	Central Anatolia (TR7)	117	5.25		
Mediterranean	Mediterranean (TR6)	270	12.12	241	10.81
Black Sea	West Black Sea (TR8)	88	3.95	176	7.90
	East Black Sea (TR9)	112	5.03		
East Anatolia	Northeast Anatolia (TRA)	120	5.39	285	12.79
	Central East Anatolia (TRB)	146	6.55		
Southeast Anatolia	Southeast Anatolia (TRC)	295	13.24	293	13.14
Total		2228	100	2229	100

The Bartlett Test of Sphericity result was found to be statistically significant ($\chi 2=28222.77$, $p<.05$); thus, it was accepted that the data were suitable for factor analysis. The result of the Kaiser-Meyer-Olkin (KMO) Test, which was carried out to test whether the sample was suitable for factor analysis, was calculated as .93. According to this statistic, it was concluded that the sample was marvelous (Hutcheson & Sofroniou, 1999) for factor analysis.

Model Specification, Estimation, and Evaluation

With the results of KMO and Bartlett Test of Sphericity, it was decided that the data were suitable for EFA. In the EFA, the Robust Maximum Likelihood estimator (as a factor extraction method, which is an extension of maximum likelihood for nonnormal continuous distributions) and GEOMIN rotation (which is one of the oblique rotation methods and allows for correlation between the latent factors) were used. First, parallel analysis (to determine the number of factors) was run with thirty-two items, and a scree plot was given in Figure 1.

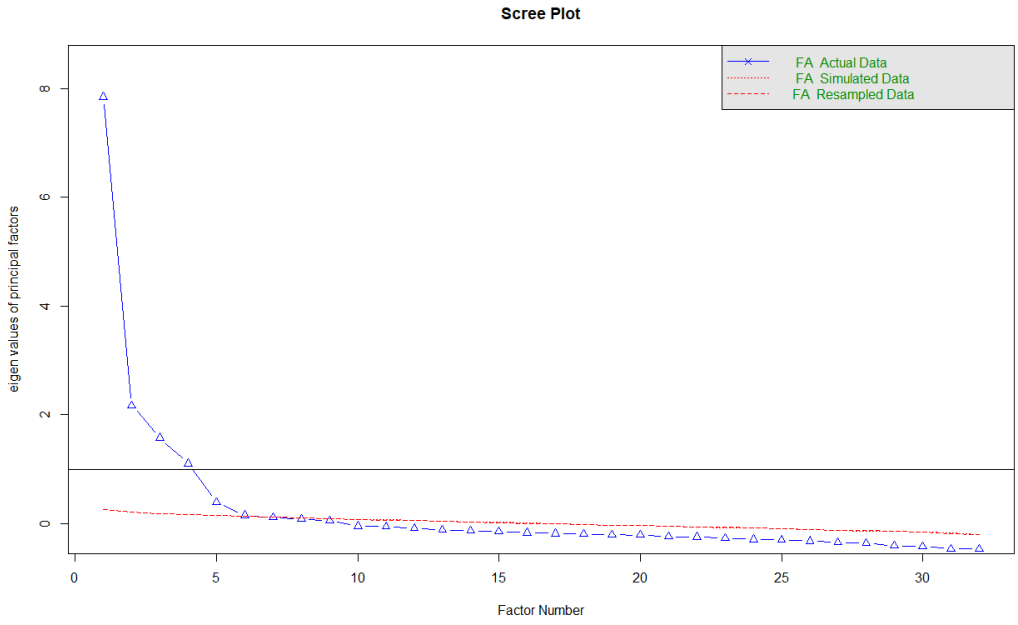

Figure 1. *Scree Plot*

When the scree plot is examined, it is seen that the elbow rule (Zoski & Jurs, 1996) is satisfied with the sixth factor. Likewise, parallel analysis suggests the number of factors as six. It is recommended not to use the Kaiser Criterion in factor retention (Howard, 2016). However, there are four eigenvalues greater than 1 and most importantly, the items' contents suggest that the construct has four factors. Considering all these, the number of factors is determined as four. After determining number of factors, EFA was run.

About the results of EFA, six items were removed due to the .40-.30-.20-factor loading rule (Howard, 2016). In this rule, it is desired for an item that a) minimum of .40 loading onto its primary factor, b) a maximum of loading .30 to another factor (s), and c) a minimum difference of .20 loading between the primary factor and other factor(s). After removing six items, EFA was re-run, and the results are given in Table 3.

Table 3. Factor Matrix

Item		Factor			
		1	2	3	4
BSBS22F	My teacher is good at explaining science	.840			
BSBS22E	My teacher has clear answers to my questions	.835			
BSBS22I	My teacher tells me how to do better when I make a mistake	.801			
BSBS22J	My teacher listens to what I have to say	.774			
BSBS22B	My teacher is easy to understand	.723			
BSBS22G	My teacher lets me show what I have learned	.701			
BSBS22H	My teacher does a variety of things to help us learn	.633			
BSBS22C	I am interested in what my teacher says	.597			
BSBS22A	I know what my teachers expects me to do	.429	.174		
BSBS21E	I like science		.867		
BSBS21I	Science is one of my favourite subjects		.834		
BSBS21A	I enjoy learning science		.809		
BSBS21F	I look forward to learning science in school		.708		
BSBS21C	Science is boring		.595		-.118
BSBS21B	I wish I did not have to study science		.505		-.131

BSBG15F	I am proud to go to this school		.662	
BSBG15C	I feel like I belong at this school		.618	
BSBG15A	I like being in school	.134	.613	
BSBG15B	I feel safe when I am at school	-.103	.594	
BSBG15G	I learn a lot in school		.520	
BSBG16C	Spread lies about me			.654
BSBG16G	Shared embarrassing information about me			.635
BSBG16E	Hit or hurt me (e.g., shoving, hitting, kicking)		-.119	.563
BSBG16B	Left me out of their games or activities			.550
BSBG16A	Made fun of me or called me names			.516
BSBG16D	Stole something from me			.482

With regard Table 3, six items (BSBG15A, BSBG15B, BSBG16E, BSBS21C, BSBS21B, and BSBS22A) loaded onto the other factors. However, no action was taken on the items since the difference in factor loadings between the factors was lower than .20, and the relatively large factor loading was in the primary factor. Based on these results, there are nine items in the first factor, and the factor loadings of items vary between .43 and .84. There are six items in the second factor, and factor loadings vary between .51 and .87. There are five items in the third factor. Factor loadings of items vary between .52 and .66. Lastly, there are six items in the fourth factor. The factor loadings vary between .48 and .66. Considering the items; these factors were named as opinions about the teacher (OT), opinions about the science lesson (OL), opinions about the school (OS), and peer bullying (PB). The total variance explained by the model, which consists of twenty-six items and four factors, is 47%.

After the factors were named, CFA was conducted. The maximum likelihood robust (MLR) method was used for estimation, as the data were continuous and had a nonnormal distribution. CFA was performed using lavaan (Rosseel, 2012) package in the R software (R Core Team, 2022). The fit values are given in Table 4.

Table 4. Fit Values of the Model

$\chi 2$	RMSEA	SRMR	GFI	CFI	TLI
$p<.05$	.051	.041	.937	.935	.928

Considering the fit statistics of CFA, RMSEA (Root Mean Square Error of Approximation; which is equal to or smaller than .05) and SRMR (Standardized Root Mean Square Residual; which is smaller than .05) values are good; GFI (Goodness-of-Fit Index; which is between .90 and .95), CFI (Comparative Fit Index; which is between .90 and .95) and TLI (Tucker-Lewis Index; which is between .90 and .95) values are acceptable. However, it was observed that the $\chi 2$ value was significant ($p<.05$). This can be explained by the fact that the $\chi 2$ value is affected by the sample size and tends to give significant results in large samples (Kline, 2011). Considering that the number of people included in the analysis was 2229, these results were interpreted as there is no issue with the $\chi 2$ value. By looking at all the fit statistics, it was concluded that the fit was sufficient. The path diagram is shown in Figure 2.

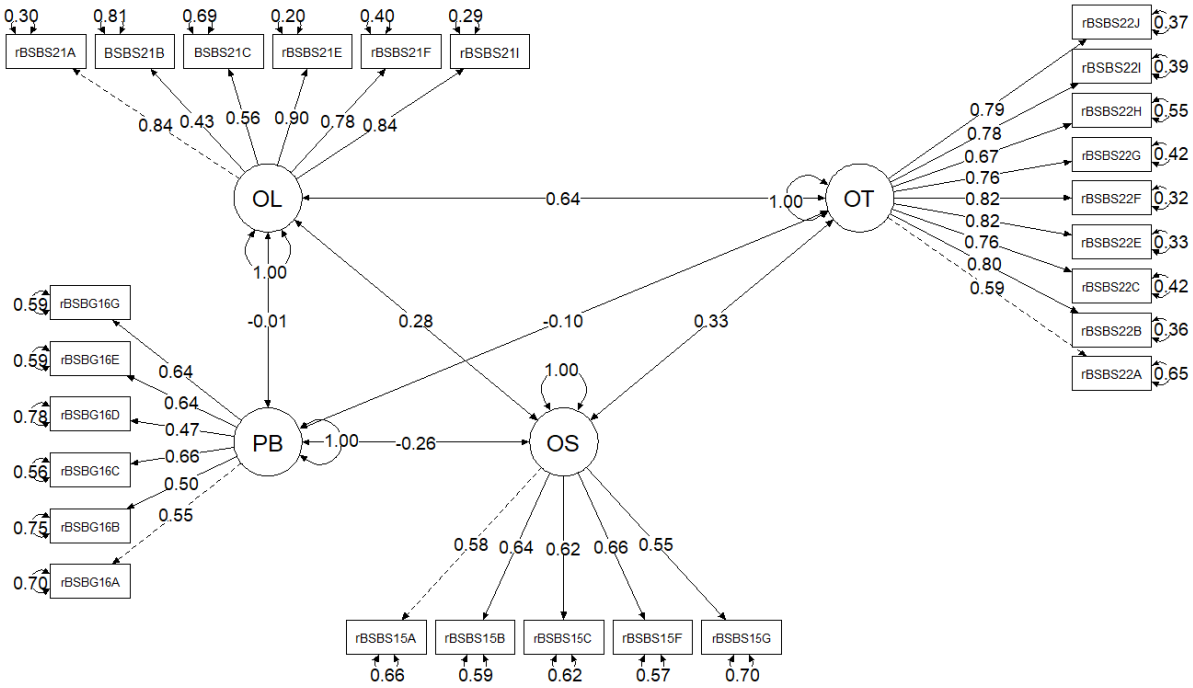

Figure 2. *Path Diagram*

When the path diagram is examined, it is seen that twenty-six observed variables in the model are grouped under four latent variables. Standardized factor loadings of the observed variables ranged from .43 to .90, and all factor loadings were significant at the p=.05 level. Error variances range from .20 to .81. When the relations between the factors were examined, a positive correlation was observed between the factors, excluding peer bullying, ranging from .28 to .64. When the peer bullying factor was examined, a relationship of -.26 was observed with attitudes towards the school. The relationship between peer bullying and attitude toward school has been demonstrated by previous studies (Çalışkan et al., 2019). However, the relationships between the teacher and liking the lesson are very close to 0. This can be explained by the fact that bullying is related to attitudes towards the school rather than the lesson or the teacher.

As a result, TIMSS 2015 Turkey eighth-grade science affective trait model was confirmed. After that, MGCFA was conducted to test the measurement invariance of the model between gender and regions.

To answer the "Does TIMSS 2015 eighth-grade science affective trait model provide measurement invariance between genders?" MGCFA was conducted using the data of male and female students simultaneously. The model, which was confirmed with all data before examining measurement invariance, was confirmed for both groups. For this purpose, CFA was conducted separately for males and females. Model fit statistics regarding gender are presented in Table 5.

Table 5. Model Fit Indices Across Genders

Gender	χ2	df	RMSEA	CFI	IFI	TLI
Female	1047.294	293	.048	.938	.938	.932
Male	1275.425	293	.055	.929	.930	.922

When the results regarding the goodness of fit in Table 5 are examined, RMSEA values are .048 and .055; CFI values of .938 and .929; IFI values are .938 and .930; TLI values were found to be .932 and .922. All the fit statistics are in the acceptable fit range. Since all fit index values were in the ideal range, it was accepted that the model had a good fit for females and males. After the confirmation of the model, MGCFA was conducted. The model fit statistics are given in Table 6.

Table 6. Measurement Invariance Models Across Genders

Invariance Model	χ2	df	RMSEA	CFI	IFI	TLI	ΔCFI
Configural	2322.719	586	.052	.934	.934	.926	-
Metric	2460.971	608	.052	.929	.929	.924	.005
Scalar	2606.154	630	.053	.924	.924	.922	.005
Strict	2853.391	656	.055	.916	.916	.917	.008

When the fit values of configural invariance shown in Table 6 are examined, the RMSEA (.052), CFI (.934), IFI (.934), TLI (.926) values were found to be in the acceptable range of fit. However, CFI, IFI, and TLI values were very close to the lower limit of the good fit interval. Since all values were at the least acceptable level, this result was interpreted as the structure is similar in all groups; that is, the model met configural invariance across gender.

When the fit values of metric invariance were examined, RMSEA (.052), CFI (.929), IFI (.929), and TLI (.924) values were seen. These results indicate that the model fits sufficiently. However, due to the possibility that χ2 statistics may give biased results in large samples, in addition to these values, ΔCFI, which shows the difference with the CFI value in the previous invariance step, was calculated. The calculated ΔCFI = .005 value was observed to be within the range of ±.01, the accepted limit. With these results, it was accepted that the factor loadings in the model were equal for male and female students; that is, metric invariance was achieved.

When the fit values of scalar invariance were examined, RMSEA (.053), CFI (.924), IFI (.924), and TLI (.922) values were seen. The results indicate that the model has also a good fit at this step. It was observed that ΔCFI = .005, which indicates the difference with the CFI value calculated in the metric invariance step, was within the range of ±.01. With these results, the item intercepts were equal for females and males, that is, metric invariance was achieved. The fact that the item intercepts are similar indicates that the gender difference in the scores has no other source other than the latent variable, that is, there is no bias at the item level. The fact that this equality is achieved makes comparing gender mean scores meaningful.

Lastly, when the fit values of strict invariance were examined, RMSEA (.055), CFI (.916), IFI (.916), and TLI (.917) values were observed. All the fit index values are in the acceptable fit range. Also, it was observed that the ΔCFI=.008 value was within the range of ±.01. It can be stated that the model meets the strict invariance.

These results show that the eighth-grade science affective trait model is invariant according to gender. Since the strict invariance condition is met, there is no problematic issue in all comparisons, including item error variances, for the affective trait model of male and female students. Considering the definition of validity, strict invariance has provided evidence for the validity of scores. According to the situation, the scores in the affective trait model can be compared between male and female students because the theoretical structure of the model is the same in both groups. In other words, the observed variable scores of females and males with equal latent variable scores are also equal.

To answer the "Does TIMSS 2015 eighth-grade science affective trait model provide measurement invariance between regions?" MGCFA was conducted by using the data of all regions simultaneously. Before the examination of measurement invariance, the model was confirmed for all regions by conducting CFA separately for all regions. Model fit statistics regarding regions are presented in Table 7.

Table 7. Model Fit Indices Across Region

Regions	χ2	df	RMSEA	CFI	IFI	TLI
Marmara	808.865	293	.052	.932	.933	.925
Aegean	538.519	293	.053	.935	.935	.928
Central Anatolia	591.110	293	.060	.903	.904	.893
Mediterranean	579.385	293	.064	.893	.894	.881
Black Sea	553.337	293	.071	.903	.904	.892
East Anatolia	574.293	293	.058	.919	.920	.910
Southeast Anatolia	660.191	293	.065	.893	.894	.881

When the results regarding the goodness of fit in Table 7 are examined, RMSEA values are .052 – .071, CFI values are .893 – .935, IFI values are .894 – .935, TLI values were found to be in the range of .881 – .928. Accordingly, RMSEA values of all regions have acceptable values. CFI, IFI, and TLI values of all regions except

Central Anatolia, the Mediterranean, the Black Sea, and Southeast Anatolia are in the acceptable fit range. However, fit values of the four regions are slightly lower than the .90 cutoff. Thus, it was accepted that the model had good fit in each region. After the confirmation of the model, MGCFA was performed. The model fit statistics are given in Table 8.

Table 8. Measurement Invariance Models Across Regions

Invariance Model	χ2	df	RMSEA	CFI	IFI	TLI	ΔCFI
Configural	4305.701	2051	.059	.916	.916	.906	-
Metric	4546.928	2183	.058	.912	.912	.908	.004
Scalar	4790.697	2315	.058	.907	.907	.909	.005
Strict	5421.270	2471	.061	.890	.889	.898	.017

When the fit values of the configural model shown in Table 8 are examined, all the values were in the acceptable range of fit. Since all values were at the least acceptable level, it was confirmed that the structure was similar in all regions.

When the fit values of the metric model were examined, all the values were in the acceptable range of fit, again. These results indicate that the model fits sufficiently. Also, ΔCFI was .004, and the value was within the range of ±.01. With these results, it was accepted that in addition to the structure in the science affective feature model, the item factor loadings in the model were the same in all regions, that is, metric invariance was achieved.

When the fit values of the scalar model were examined, all the values were in the acceptable range of fit, again. These results indicate that the model has a good fit at this step. Also, ΔCFI=.005 value was seen. Thus, it was accepted that the item intercepts were the same in all regions; metric invariance was achieved. The fact that the item intercepts are similar indicates that the regional difference in the scores has no source other than the latent variable; that is, there is no bias at the item level. The fact that this equality is achieved makes the comparison of region mean scores meaningful.

Lastly, when the fit values of the strict model were examined, RMSEA (.015) was in the acceptable range of fit, but CFI (.890), IFI (.889), and TLI (.889) values were seen. Although the RMSEA value was acceptable and CFI, IFI, and TLI were very close to the acceptable limit, it was observed that the ΔCFI=.017 value, which was not within the range of ±.01. This shows that the error variances were not similar in all groups; the science affective trait model did not provide strict invariance between regions.

The results show that the science affective trait model is only partially invariant across regions. Since the scalar invariance is ensured, correct interpretations can be made when the affective trait scores are compared according to the regions. However, comparing the error variances will not be correct since there is no full invariance. The lack of strict invariance can be explained by the differences between the regions included in this study, although strict invariance is a difficult condition. The development level, population density, geographical characteristics of the regions, and the education level of the parents of the students residing in the regions, and socioeconomic status are distinctly different. The fact that the regions are significantly different indirectly affects the difference between affective trait scores.

DISCUSSION, CONCLUSION & RECOMMENDATIONS

In this study, the measurement invariance of the science affective trait model, created using the items in the TIMSS 2015 eighth-grade science student questionnaire, was tested according to gender and regions in NUTS Level 1 in Turkey. As a result of the analysis, it was seen that the model met all the invariance steps for gender and the scalar invariance condition for the regions but not the strict invariance.

The fact that the configural invariance was met shows that the model is equal for males and females. Thus, the model measures the same structure for them. Achieving metric invariance shows that the factor loadings of the items are equal for male and female students. The equality of factor loading means that male and female students do not interpret the items in the model differently from each other. Meeting the scalar invariance shows that the item intercepts in the model are equal for male and female students, the items do not show any bias in favor of females or males, and the source of the difference in the responses to the items is the difference in the latent variable scores. In this case, the observed mean of variables of female and male students can be compared. The strict invariance step examines whether the error variances are the same between the groups. However, as the latent variable variances increase, the error variances of the observed variables also increase, so strict invariance is a

complex condition to meet in practice. The fact that strict invariance was confirmed for gender shows that the error variances in the model are similar for males and females. The variance and covariance of the observed variable scores are suitable for comparison by gender. Uzun (2008) established a model with the variables of attitude towards the course, the importance given to the course, self-efficacy, and student activities for the classroom that affected science achievement in the TIMSS-R Turkey application and tested the measurement invariance of this model in terms of gender. According to the results, while the attitude factor met scalar invariance, the other three factors met the metric invariance. Although the results of the two studies are partially similar, the different results can be explained by the differences in the variables in the models.

Since configural invariance is ensured, the model represents the same structure in all regions. Thus, the structure can be compared across regions. Meeting the metric invariance means that the factor loadings of the model are equal for all regions. According to this result, students from all regions similarly interpreted the items. Achieving scalar invariance means item intercepts are equal in all regions and show no bias in any region's favor. According to this result, comparing the means of the items in the model according to the regions does not lead to any erroneous interpretation. Although the strict invariance is necessary for a fair and unbiased comparison between groups, it is difficult to achieve because the error variances are directly proportional to the latent variable. Failure to meet strict invariance between regions means that the error variances in the model are not equal in all regions. Thus, it is not possible to compare the observed variances and covariances by regions meaningfully. Ölçüoğlu and Çetin (2016) established a three-factor model that affects eighth-grade mathematics achievement in Turkey with TIMSS 2011 data and tested the measurement invariance of this model according to seven geographical regions in Turkey. As a result of the analysis, it was stated that the model met scalar invariance but did not meet strict invariance. Polat (2019) created a four-factor model that affected eighth-grade science achievement in Turkey with the data of TIMSS 2015 and examined the measurement invariance of this model according to the regions in Turkey NUTS Level 1. It has been observed that the model meets scalar invariance between regions but did not meet the strict invariance. The results obtained from these studies are similar to those of the studies in the literature.

While the model is entirely invariant between gender, it is not between regions. The lack of complete invariance between regions affects the validity negatively. In this case, no evidence was provided for comments on the affective trait scores of the regions. This result should be considered when comparing regions considering the effect of affective characteristics on course achievement. Full invariance should be achieved between regions. In future research, the characteristics of the individuals to be studied should be considered. Invariance studies should be carried out to find solutions to the problems that may arise in terms of validity in the beginning.

It has been observed that the factor of opinions about the teacher has the most significant positive effect. When we look at the items with the most significant factor loadings in this factor, the items related to the teacher's power of expression in the course and explaining the mistakes are seen. Improvement in teachers' ability to express themselves will positively affect students' science achievement. Possible studies to increase science achievement will provide significant gains if there are studies to improve the expressive power of teachers.

In this study, measurement invariance analysis was carried out at the gender and inter-regional level using the MGCFA method with the data of the items that were thought to affect the science course achievement in the eighth-grade student questionnaire in the TIMSS 2015 Turkey application. Comparisons can be made between the two grade levels by using the fourth-grade data and the eighth-grade or the two models by creating a model for the mathematics and science courses. In addition to the variables used in the model, other variables can be used to test measurement invariance through a more comprehensive model. Also, in addition to the student questionnaire data, a separate model can be established for each questionnaire using teacher and school administrator questionnaires, or a more comprehensive research model can be established by analyzing all the questionnaire data simultaneously with the multi-level SEM. In addition to the MGCFA method, studies comparing the results of different methods can be carried out using other methods in which measurement invariance is examined.

Statements of Publication Ethics

This research was reviewed by the Ethics Commission of Hacettepe University Senate, and it was approved that the research was ethically appropriate. Meeting date and ethical decision number: 18/06/2018, 35853172-300.

Researchers' Contribution Rate

Both authors equally contributed to the literature review, method, data collection, data analysis, results, discussion, and conclusion sections.

Conflict of Interest

The authors reported no potential conflict of interest.

Acknowledgement

The present study is part of a master thesis conducted under the supervision of the second author and prepared by the first author. This study acknowledges the significance of the role of the Ministry of National Education in providing data of regional codes.

REFERENCES

Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6*(1), 18-47.

Bahadır, E. (2012). *Uluslararası öğrenci değerlendirme programına (PISA 2009) göre Türkiye'deki öğrencilerin okuma becerilerini etkileyen değişkenlerin bölgelere göre incelenmesi [According Programme for International Student Assessment (PISA 2009), investigation of variables that affect Turkish students' reading skills by regions]* (Unpublished master's thesis). Hacettepe University Institute of Social Sciences.

Başusta, N. B. (2010). Ölçme eşdeğerliği [Measurement equivalence]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(2), 58-64.

Başusta, N. B. & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire]. *H. U. Journal of Education, 30*(4), 80-90.

Bialosiewicz, S., Murphy, K., & Berry, T. (October, 2013). *An introduction to measurement invariance testing: Resource packet for participants. Do our measures measure up? The critical role of measurement invariance?* Demonstration Session American Evaluation Association, Washington, DC Retrieved from http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Bloom, B. (2012). *İnsan nitelikleri ve okulda öğrenme [Human qualities and learning in school]. (*D. A. Özçelik, Ed. & Trans.). MEB Publishing. (Original work published 1979).

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley-Interscience Publication.

Bryne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175. https://doi.org/10.1177%2F0022022102250225

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. & Demirel, F. (2012). *Bilimsel araştırma yöntemleri [Scientific research methods]* (12th edition). Pegem Akademi Publishing.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://psycnet.apa.org/doi/10.1037/h0040957

Çalışkan, Z., Evgin, D., Bayat, M., Caner, N., Kaplan, B., Öztürk, A., & Keklik, D. (2019). Peer bullying in the preadolescent stage: frequency and types of bullying and the affecting factors. *Journal of Pediatric Research, 6*(3), 169-179. https://10.4274/jpr.galenos.2018.26576

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

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it [Editorial]. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups?: Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(1), 78-94. https://dx.doi.org/10.1097%2F01.mlr.0000245454.12228.8f

Gülleroğlu, H. D. (2017). PISA 2012 Matematik uygulamasına katılan Türk öğrencilerin duyuşsal özelliklerinin cinsiyete göre ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance by gender for the Turkish students' affective characteristics who took the PISA 2012 math test]. *Gazi University Journal of Gazi Educational Faculty (GUJGEF), 37*(1), 151-175.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51-62. https://doi.org/10.1080/10447318.2015.1087664

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist*. Sage.

Karakoç Alatlı, B., & Çokluk Bökeoğlu, Ö. (2018). Uluslararası Öğrenci Değerlendirme Programı (PISA-2012) okuryazarlık testlerinin ölçme değişmezliğinin incelenmesi [Investigation of Measurement Invariance of Literacy Tests in the Programme for International Student Assessment (PISA-2012)]. *Elementary Education Online, 17*(2): pp. 1096-1115. http://dx.doi.org/10.17051/io.2015.85927

Kıbrısoğlu, N. (2015). *PISA 2012 matematik öğrenme modelinin kültürlere ve cinsiyete göre ölçme değişmezliğinin incelenmesi: Türkiye – Çin (Şangay) – Endonezya örneği* [The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey – China (Shangai) - Indonesia] (Unpublished master's thesis). Hacettepe University Institute of Educational Sciences.

Kline, R. B. (2011). *Principles and practices of structural equation modeling*. The Guilford Press.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388. https://doi.org/10.1177%2F1094428104268027

MoNE (2016). *TIMSS 2015 Ulusal Matematik ve Fen Ön Raporu: 4. ve 8. Sınıflar [TIMSS 2015 National Mathematics and Science Preliminary Report: 4th and 8th grades]*. Retrieved from https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/23161945_timss_2015_on_raporu.pdf

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380-392). Guilford Press.

Mullis, I. V. S., & Martin, M. O. (Eds.) (2013). *TIMSS 2015 Assessment Frameworks*. Retrieved from https://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Mathematics.pdf

Öğretmen, T. (2006). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği [he investigation of psychometric properties of the test of progress in international reading literacy (PIRLS) 2001: The model of Turkey-United States of America]* (Unpublished doctoral dissertation). Hacettepe University Institute of Social Sciences.

Ölçüoğlu, R., & Çetin, S. (2016). TIMSS 2011 sekizinci sınıf matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi [The investigation of the variables that affecting TIMSS 2011 eight grade math achievement according to regions]. *Journal of Measurement and Evaluation in Education and Psychology, 7*(1), 202-220.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529. https://psycnet.apa.org/doi/10.1037/0021-9010.87.3.517

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. https://psycnet.apa.org/doi/10.1037/0033-2909.114.3.552

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implication. *Human Resources Management Review, 18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (5th Edition). Pearson Education.

Uyar, Ş. (2011). *PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample]* (Unpublished master's thesis). Hacettepe University Institute of Social Sciences.

Uzun, N. B. (2008). *TIMSS-R Türkiye örnekleminde fen başarısını etkileyen değişkenlerin cinsiyetler arası değişmezliğinin değerlendirilmesi [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey sample]* (Unpublished master's thesis). Hacettepe University Institute of Social Sciences.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177%2F109442810031002

Widaman, K. F., & Reise, S. P. (1997). *Exploring the measurement invariance of psychological instruments: Applications in the substance use domain*. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 281–324).

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3). https://doi.org/10.7275/mhqa-cd89

Yandı, A., Köse, İ. A., Uysal, Ö. & Oğul, G. V. (2017). *PISA 2015 öğrenci anketinin (St094Q01na - St094Q05na) ölçme değişmezliğinin farklı yöntemlerle incelenmesi [Investigation of measurement invariance of PISA 2015 student questionnaire (St094Q01na - St094Q05na) with different methods]*. In Demirel, Ö. and Dinçer, S. (Ed.), Küresel Dünyada Eğitim, (pp. 333-344). Pegem Akademi Publishing.

Zoski, K.W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. *Educational and Psychological Measurement*, *56*, 443–451. https://doi.org/10.1177/0013164496056003006