

## Investigation of Measurement Invariance of Science Motivation and Self-Efficacy Model: PISA 2015 Turkey Sample

Metehan Gungor <sup>1,\*</sup>, Kubra Atalay Kabasakal <sup>2</sup>

<sup>1</sup> Ministry of National Education, 06450, Ankara, Turkey

<sup>2</sup> Hacettepe University, Education Faculty, Measurement and Evaluation Department, 06800, Ankara, Turkey

### ARTICLE HISTORY

Received: Oct 16, 2019

Revised: Apr 10, 2020

Accepted: May 01, 2020

### KEYWORDS

Structural equation modeling,  
Measurement invariance,  
Instrumental motivation,  
Science self-efficacy

**Abstract:** Measurement invariance analyses are carried out in order to find evidence for the structural validity of the measurement tools used in the field of educational sciences and psychology. The purpose of this research is to examine the measurement invariance of Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales found in the PISA 2015 Student Questionnaire across different groups in the Turkish sample survey. The analysis was carried out with the data obtained from 4583 students that met the analysis assumptions. The measurement invariance of the model in terms of gender and statistical regional groups was examined by the structural equation modeling (SEM) technique. Firstly, the data was examined to determine whether the assumptions for the analyses were met. Then, measurement models were verified by performing confirmatory factor analysis (CFA). The measurement invariance across genders and statistical regions was tested by multi-group confirmatory factor analysis (MGCFA). Unweighted Least Squares (ULS) method was used as the estimation method in CFA and MGCFA stages. In order to make final decisions about the stage of measurement invariance models hold, Comparative Fit Index (CFI) was used. The results of the study show that the research model ensures all stages of invariance across gender groups and regions. Science Motivation and Self-Efficacy Model illustrates that it is valid to make comparisons between scores of male and female students or students from different regions of Turkey. According to the findings, the research model could provide complete measurement invariance.

## 1. INTRODUCTION

Education indicators (budget allocated for education, using education technologies, quality of people working within the field of education, literacy rate, etc.) provide significant information as to the development level of countries. Performances of students that are included in a specific education system can be accepted to be a good indicator of the quality of this education system in a country. Measurement and evaluation instruments are frequently used while making comparison related to the performances of students. Implementing measurement and evaluation at the national and international level plays an important role in developing educational policies of countries. There are a great number of assessment and evaluation implementations around the world. The most popular ones include large scale exams such as Programme for

CONTACT: Metehan Gungör ✉ [gungormetehan@gmail.com](mailto:gungormetehan@gmail.com) 📍 Ministry of National Education, 06450, Çankaya/Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). PISA is a large-scale triennial assessment and it focuses on how 15-year-old students make use of their already-existing knowledge and skills to solve daily life problems. Each implementation is grounded on one of the fields among reading skills, mathematical literacy and science literacy. While the main subject field of PISA was reading skills in 2000, which is the year when PISA was implemented for the first time, the conceptual foundation of the sixth cycle in 2015 was science literacy. PISA is a useful assessment instrument in terms of evaluating the effectiveness of education systems via collecting data on the basis of students, teachers and schools, and using the results obtained at the end of analyzing these data. Taking advantage of PISA to monitor the dispositions in the knowledge and skills of students coming from different countries and different demographic regions of each participant country leads the drive for developing knowledge (Organisation for Economic Co-Operation Development [OECD], 2017). PISA can be regarded as a watershed in the discourse on education in many countries. Countries make use of the results of PISA while they are developing their own educational policies. Some new practices on assessment and curriculum standards have been reformed and PISA-like competencies have been incorporated into their systems. Therefore, it is important to stay careful when interpreting PISA results regarding comparability. It should be tested whether sub-scales that are used in assessment measure the same construct in each sub-group. Otherwise, the interpretation that is based on the results of the assessment will not be valid. Construct equivalence is a basic assumption that should be met if the developers or executors of any assessment aim at comparing the scores of different groups or interpreting these scores in compliance with the intended use (Gierl, 2000). When an assessment instrument is designed to compare participants coming from two or more cultures, the construct to be measured via the test should be equivalent for the comparison to be meaningful (Hambleton, 1994). The necessity to examine if the structures to be measured via tests are equivalent or not makes the issue of measurement invariance a leading topic within the scope of assessment and evaluation implementations.

### **1.1. Measurement Invariance**

Measurement invariance is described by Byrne and Watkins (2003) as “the level of items being perceived and interpreted the same among groups.” On the other hand, Mellenbergh (1989) as well as Meredith and Millsap (1992) starts out from the concept of “biasness” and describes measurement invariance as “the state of the conditional probability of obtaining a specific observed score related to an ability being independent of group membership in mathematical terms.” In other words, measurement invariance is measuring a psychological construct with the same level of correctness in all sub-groups (Sireci, Patsula, & Hambleton, 2005). Measurement invariance is a special property that should be tested in order for the between-group comparisons of the psychological construct that will be measured to be meaningful (Cheung & Rensvold, 2002) and for the deductions and interpretations that will be made at the end of comparisons to be valid (Somer, Korkmaz, Dural, & Can, 2009). Measurement invariance analysis can be carried out to find proof for the structural validity of tests that are developed to draw between-group comparisons. The test of measurement invariance is a kind of covariance structure analysis and it is designed on the basis of measuring a specific structure on different groups (Başusta, 2010). The most common method of testing measurement invariance is Multi-Group Confirmatory Factor Analysis (MG-CFA) that falls under the umbrella term of Structural Equation Modeling (SEM) (Jöreskog & Sörbom, 1999; Kline, 2011; Koh & Zumbo, 2008). While SEM includes measurement errors in the model, it also considers the direct and indirect effects of the variances in the created model. Hence, it makes it possible to test, estimate and develop multivariate complex models (Raykov & Marcoulides, 2006). Four hierarchical nested models should be tested while examining measurement invariance with

MGCFA under the umbrella term of SEM. These four hierarchical models can be listed as (1) configural invariance, (2) weak invariance, (3) strong invariance and (4) strict invariance respectively (Meredith, 1993; Wu, Li & Zumbo, 2007).

### 1.1.1. *Configural invariance*

This is the first step of measurement invariance test. The groups are tested to see if they have the same factor structure or not at this stage. For that purpose, equivalence of factors and pattern of factor loading is analysed at this stage (Taris, Bok & Meijer, 1998). If configural invariance is ensured at the end of the analysis, this means that the same structure is measured in the comparison groups. If the analysis shows that the conditions of configural invariance are not met, this means that different structures are measured among groups (Wu, Li & Zumbo, 2007). Kline (2011) states that if the necessary conditions are not met at this stage, measurement invariance cannot be ensured at more constrained stages.

### 1.1.2. *Weak invariance*

The equivalence of measurement unit or factor loadings are analysed at this stage. It is tested if the groups have the same measurement unit concerning the latent variable or not at this stage of weak invariance. Therefore, this stage, which can be described as the test of the measurement unit, is called metric invariance. In this model, factor loadings are also restricted in addition to the conditions that are valid at the stage of configural invariance (Vandenberg & Lance, 2000). If the weak invariance cannot be ensured, it can be discussed that factors do not mean the same in different groups (Gregorich, 2006).

### 1.1.3. *Strong invariance*

It is tested if the constant of regression that is obtained when the factor scores of the groups to be compared is zero is equal or not at this stage of strong invariance. Because of this reason, strong invariance is also called as scalar invariance (Vandenberg & Lance, 2000). The equivalence of the observed variables and factor loadings are also examined at this stage, which requires between-group equivalence of factor variance and covariances. If the necessary conditions are met, it means that the means of the observed variables and factor loadings can be compared (Gregorich, 2006).

### 1.1.4. *Strict invariance*

While invariance is tested, parameter restrictions as well as error variances are limited at this stage (Vandenberg & Lance, 2000; Wu, Li & Zumbo, 2007). This is the last step of measurement invariance test. Ensuring this stage is proof of measurement invariance. Assessment tools that claim to be measuring the same construct among the groups should meet the conditions of strict invariance. Measurement invariance can be ensured only if this stage is ensured. The stages and the related conditions in question are summarized in [Table 1](#).

**Table 1.** *Measurement invariance stages.*

Degree of Invariance	Condition of Invariance	Group Comparison
Configural Invariance	Item/Factor groups	-
Weak Invariance	Item/Factor groups and factor loadings	Factor variance and covariances
Strong Invariance	Item/Factor groups, factor loadings and item intercepts	Factor variance and covariances, factor and observed variable averages
Strict Invariance	Item/Factor groups, factor loadings, item intercepts, and item residual variance	Factor variance and covariances, factor and observed variable averages, observed variance and covariances

[Kıbrıslıoğlu Uysal & Akın Arıkan, 2018]

## 1.2. Instrumental Motivation to Learn Science

Motivation is a psychological construct that affects student success and it provides people with the necessary power to carry out a specific activity (Schunk, Meece & Pintrich, 2014). When motivation is addressed within the scope of learning, it is the power that stimulates, maintains and directs the behaviour towards a specific goal (Dilts, 1998). If the student thinks that the knowledge that s/he has acquired in a lesson will be useful in her/his life and career, s/he can make a great effort in this lesson even if the topics in the lessons are not interesting for her/him (İlhan, 2015). This effort is influential on this student's performance. Such a motivation is called instrumental motivation. Instrumental motivation means that students discern that what they have learnt will be useful in their future studies and career plans, and so they are eager to learn science (Wigfield & Eccles, 2000). Student motivation is an indispensable part of a qualified educational life. Although it is widely accepted that motivation is an indispensable part of education, it is not known well how to use motivation in instructional design and what it means in this context. This results from the fact that motivation is a construct. Instrumental motivation in learning science under the heading of motivation to learn science was measured via a four-point Likert type sub-scale consisting of four items in the cycle of 2015. The items included in the sub-scale try to measure if the students think that science lesson will be useful in their future educational life and career plans (OECD, 2016). The items included in the Instrumental Motivation to Learn Science Scale, which is the subject matter of this study, is given in Table 2.

**Table 2.** *Items that constitute Instrumental Motivation to Learn Science Scale in PISA 2015 student questionnaire.*

Code	Item
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on.
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on.
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.

## 1.3. Science Self-Efficacy

Efficacy belief is one of the concepts that underpin the Social Learning Theory developed by Bandura. Bandura (1997) describes self-efficacy as the judgments of individuals as to what they are able to do and their belief as to the ability to fulfill a specific task successfully or display a behaviour accomplished. Self-efficacy is individuals' own judgments about how well they are able to fulfil an activity that is necessary for the solution to possible problems (Bıkmaz, 2002). Self-efficacy is not related to people's knowing what to do, but to their belief about what they are able to do or learn (Schunk & Pajares, 2009). With reference to these descriptions, it seems possible to describe science self-efficacy as individuals' own judgments about how well they are able to do the specific tasks that are necessary for finding solutions for problems in science. Science self-efficacy is one of the fields on which many studies focus on the literature (Bakan Kalaycıoğlu, 2015; Bircan & Sungur, 2016; Britner & Pajares, 2001; Uzun, Gelbal & Öğretmen, 2010; Zedlin, Britner & Pajares, 2007). It was concluded at the end of a study carried out by Uzun, Gelbal and Öğretmen (2010) in order to examine the relationship between success in science and cognitive qualities that science self-efficacy is the most important variable used to explain the male and female students' success in science. Science self-efficacy was measured in PISA in 2006 and 2015 with the same four-point Likert-type sub-scale that consisted of eight items (OECD, 2016). Whereas PISA is implemented in the same subject field once every nine

years, similar and same items are used for students. The purpose of using some common items in these implementations is to examine the trends in education. The items that constitute Science Self-Efficacy Scale, which is the subject matter of this study, are given in [Table 3](#).

**Table 3.** *Items that constitute Science Self-Efficacy Scale in PISA 2015 student questionnaire.*

Code	Item
ST129Q01TA	Recognise the science question that underlies a newspaper report on a health issue.
ST129Q02TA	Explain why earthquakes occur more frequently in some areas than in others.
ST129Q03TA	Describe the role of antibiotics in the treatment of disease.
ST129Q04TA	Identify the science question associated with the disposal of garbage.
ST129Q05TA	Predict how changes to an environment will affect the survival of certain species.
ST129Q06TA	Interpret the scientific information provided on the labelling of food items.
ST129Q07TA	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.
ST129Q08TA	Identify the better of two explanations for the formation of acid rain.

#### **1.4. Literature Review**

The interpretations of the scores that are obtained from the measurement tools may vary among different groups. If the scores obtained from the same test are not comparable among different groups (gender, culture, socio-economic level, etc.), the differences in the mean scores of the groups or the correlation patterns of the test with external variables are potentially artificial and they may be misleading to a great extent (Reise, Widaman & Pugh, 1993). A test that is implemented with different groups may ensure measurement invariance while it does not serve for the measurement invariance among genders. This may result from between-group real differences whereas it may also result from the assessment tool itself (Başusta & Gelbal, 2015). For that reason, measurement invariance studies are carried out with large scale international tests such as PISA, TIMSS and PIRLS, whose results are preferred to make comparisons (Akyıldız, 2009; Alivernini, 2011; Asil & Gelbal, 2012; Ayvalli & Biçak, 2018; Başusta & Gelbal, 2015; Ercikan & Koh, 2005; Ertürk & Erdiñç Akan, 2018; Gülleroğlu, 2017; Karakoç Alatlı, Ayan, Polat Demir & Uzun, 2016; Kıbrıslıoğlu Uysal & Akın Arıkan, 2018; Nagengast & Marsh, 2014; Oliden & Lizaso, 2013; Ölçüoğlu & Çetin, 2016; Scherer, Nilsen & Jansen, 2016; Uyar & Doğan, 2014; Uyar & Kaya Uyanık, 2019; Wu, Li & Zumbo, 2007). For example, Ölçüoğlu and Çetin (2016) modelled some variables that affect the maths success of 8-grade students that participated in TIMSS 2011 in Turkey, and they examined the measurement invariance with MGCFA among the seven regions in Turkey. The sample of this study was composed of 6928 14-year-old students chosen from 239 schools in Turkey. The results showed that only configural invariance and weak invariance were maintained in sub-groups of regions. According to this result, the scale could not meet the conditions of invariance and strict factorial invariance cannot be detected. Therefore, according to the findings of the study, it wouldn't yield valid results to make a comparison between regions with the the scores which have been obtained via the items that constitute the subject matter of the study and that are deemed to have an effect on maths success of students. Gülleroğlu (2017) carried out a study with the data of PISA 2012 in order to examine the measurement invariance of affective qualities towards maths according to the variable of gender. The sample of the study was composed of 1598 students that took Form B in the test and were chosen from 15-year-old 4848 students in 170 schools in Turkey. The measurement invariance of factors which can be listed as interest in mathematics, anxiety for maths, self-perception towards maths and self-efficacy were examined with MGCFA. The researcher reported at the end of the study that all the variables apart from self-efficacy met the conditions of configural invariance. The researcher, who examined the measurement invariance through hierarchical measurement invariance, concluded that strict

invariance was not ensured in all the five factors that were analyzed. Therefore, strict measurement invariance could not be provided in all five factors that were the subject matter of the study. Ertürk and Erdiñç Akan (2018) carried out a study of measurement invariance according to gender with the data of TIMSS 2015. The study aimed at examining the measurement invariance of some variables related to the success of maths according to gender. The sample of the study consisted of 6456 4-grade students who participated in the test in Turkey. The latent variables that were chosen for the study were liking mathematics, interest in mathematics and confidence in mathematics, which were all thought to have an impact on maths. Each variable was examined separately and MGCFA was used to examine the measurement invariance hierarchically. The differences between the values of CFI (comparative fit index) among the invariance were tested. It was found out at the end of the study that all the variables that were tested in the study met the conditions of configural invariance whereas only the variable of liking mathematics met the conditions of strict invariance. Uyar and Dođan (2014) carried out a study with PISA 2009 Turkey sample in which they established a model called learning strategies and they examined the measurement invariance of the model according to the statistical regions in Turkey. The researchers reported that the model met the conditions of strict measurement invariance. For instance, Uyar and Kaya Uyanık (2019) established a different model with the affective scales that are also the subject matter of this study and that were included in PISA 2015, and they examined the gender-based measurement invariance with the sample of Turkey. As a result, the model that was constructed by the researchers could only meet the conditions of the stage of weak invariance. On the contrary to these studies, Ayvallı & Biçak (2018) carried out a study with one of the affective tests of PISA 2012 Turkey sample and reported that strict measurement invariance was ensured after doing analysis between genders. Similarly, Bařusta and Gelbal (2015), who carried out a study with the data of PISA 2006 implementation, reported that strict measurement invariance was ensured between gender groups. It can be concluded from the results of these studies that it is possible to have different findings as to whether similar and same tests provide measurement invariance or not.

Polat and Madra (2018) carried out a study based on gender by using the data obtained from both PISA and TIMSS 2015 and they found out at the end of the study that female students in Turkey were far behind male students about turning advantageous qualities such as self-confidence, sense of belonging to school, motivation, liking to learn into success. There are inconsistent results in the studies based on gender about the success in the field of maths and science (Ađaç & Masal, 2015; Batyra, 2017; Larson, Stephen, Bonitz & Wu, 2014). Starting from study results, the variables that are chosen for this study are Instrumental Motivation to Learn Science and Science Self-Efficacy, which are thought to have an impact on science literacy. One should be sure that the same assessment tool is used in all of the groups in order to interpret the research findings correctly. If the sub-scales included in the PISA Student Questionnaire provide measurement invariance, this means that the same qualities are measured across different groups. Only the data obtained in this way can be comparable across groups. There are a number of measurement invariance studies that use gender and regions as variables (Bařusta & Gelbal, 2015; İmrol, 2017; Kıbrıřlıođlu, 2015; Ölçüođlu & Çetin, 2016; Uyar & Dođan, 2014). Different results have been reported in these studies.

### **1.5. Aim of the Study**

Aim of the study is to examine the measurement invariance of Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales in the PISA 2015 Student Questionnaire across different groups in Turkish sample. Measurement invariance of a model that was constructed with the scales used in PISA was examined in this study. The differences between the groups to which the scales are applied

can result from the real differences between the groups, whereas they may also arise from the scales themselves. In this study, it is aimed to provide evidence of the validity of inferences based on differences between groups by investigation of comparability.

Turkey’s variation in performance between schools is particularly large and is about twice the OECD average between-school variance. Therefore, it is thought that the investigation of the comparability of the region groups is especially important. In this context, it is expected that the study results will contribute to the literature of measurement invariance.

## **2. METHOD**

### **2.1. Research Method**

This study is a descriptive study as it aims at identifying whether Science Motivation and Self-Efficacy Model constructed by Instrumental Motivation to Learn Science and Science Self-Efficacy subscales in the Student Questionnaire of PISA 2015 is invariant by gender and statistical regions of Turkey.

### **2.2. Population and Sample**

The sixth cycle of PISA, which is implemented by the Ministry of National Education, Directorate General for Measurement, Assessment and Examination Services, was carried out in 2015 in a computer-based way with the participation of 5895 students in Turkey. Population of PISA 2015 Turkey implementation was determined to be 15-year-old 1.324.089 students while the sample to be able to participate in the implementation was found to be 925.366 students (Ministry of National Education [MNE], 2016).

The sample of this study is composed of 4583 students that met the analysis assumptions. Some statistical regions have been excluded from the scope of this study in accordance with the results of CFA that was conducted for the variables of gender and statistical region separately. The information about the regions that have been excluded from the study is given in FINDINGS section. The figures of the sample with which the study was carried out are given in [Table 4](#).

**Table 4.** *Sample of the study by gender and statistical regions.*

Gender	n	%
Female	2318	50.6
Male	2265	49.4
Region	n	%
Istanbul (TR1)	837	26.3
Aegean (TR3)	562	17.6
West Anatolia (TR5)	456	14.3
Meditarrenean (TR6)	669	21
Middle East Anatolia (TRB)	179	5.6
Southeast Anatolia (TRC)	485	15.2

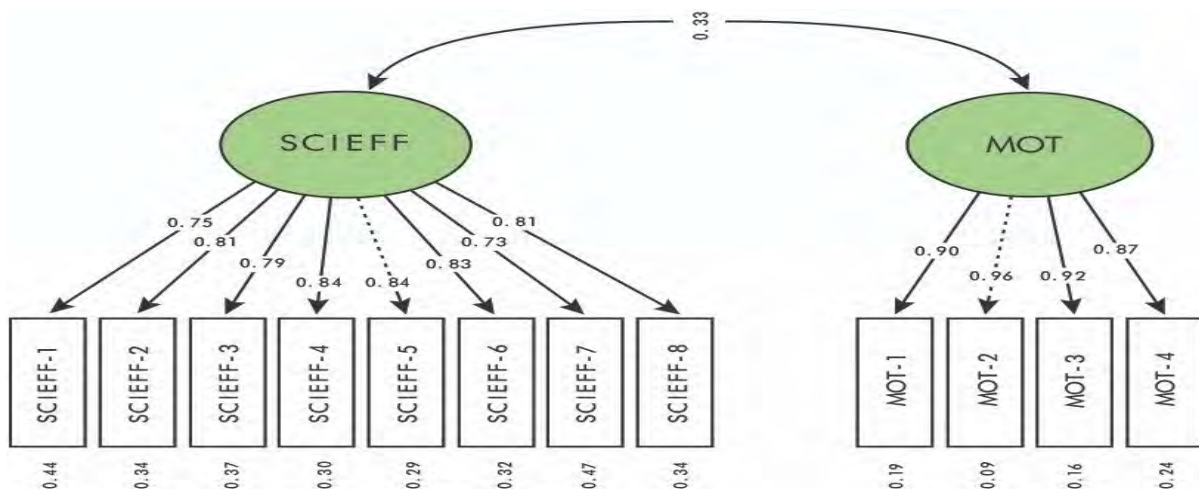
Note. TR1, TR3, TR5, TRB and TRC are the codes of the given regions.

### **2.3. Data Analysis**

Because of the reason that all multi-variable statistical techniques are based on assumptions to a certain extent (Çokluk, Şekercioğlu & Büyüköztürk, 2016), the assumptions of (1) missing values, (2) extreme values, (3) normality, (4) multicollinearity, (5) linearity, (6) homogeneity and (7) sample size are tested.

All the analysis carried out within the framework of this study were done by taking the advantage of open-source R software (R Development Core Team, 2017). The scale items which were not answered by the students were accepted to be missing values in this study. As

there was not a specific pattern among the data of students who had missing answers (the item that had the highest missing value was the item coded ST129Q03TA with 374 data, which represented 6.3% of the data set), data gathered from 1063 students were excluded from the data set. It was examined in the study whether there was a univariate extreme value or not. For that purpose, z scores of 12 items in two different scales were calculated. It was found out that all z scores were between the values of -3 and +3. Afterwards, Mahalanobis distances of the variables were examined. The critical values of  $\chi^2$  were examined when  $p < 0.001$ . The value of 32.9095 was obtained for the degree of freedom ( $df$ ) of 12 and 249 data that were above this value were excluded from the study. The values of skewness and kurtosis of the variables were taken into consideration in order to decide whether the data had a normal distribution or not. It was decided at the end of the statistical analysis that the data had normal distribution. The items that are the subject matter of this study were examined for multicollinearity analysis, and the item coded ST113Q02TA (MOT-2) had the highest VIF value of 4.107 among the items belonging to the Scale of Instrumental Motivation to Learn Science. Similarly, the item coded ST129Q05TA (SCIEFF-5) had the highest VIF value of 2.403 among the items belonging to the Scale of Science Self-Efficacy. Moreover, CI values of items were examined and the highest CI value among the items of the scale was found to be 22.594. With reference to the results of these analyses, it can be stated that there is not a problem of multicollinearity among the items used in this study (Gujarati, 1995; Kline, 2011). Scatter matrix was used for the linearity assumption in the study. It is expected for linearity that diagrams formed by the variable pairs should be ellipsis or ellipsis-like shapes, but linearity assumption could not be ensured in this study. Homoscedasticity can also be examined with Box-M test in multi-variable statistics. When Box-M test is found to be significant ( $p < 0.05$ ), it can be concluded that homoscedasticity assumption cannot be ensured. Box-M test had significant results in this study. When multi-variate normality is ensured, the relation among variables can be said to be homoscedastic (Tabachnick & Fidell, 2013). It can be stated that statistics related to linearity and homoscedasticity assumptions are not enough to ensure normality assumption. Although there are debates about the adequate size of the sample, it is stated that the smallest sample for SEM analysis should be over 150 (Anderson & Gerbing, 1988; Kline, 2005). In this study, the assumptions of missing values, extreme values, normality, multicollinearity as well as size of the sample were examined. The data obtained from 4583 students ensuring the assumptions were found to be enough for SEM analysis. In this study, CFA was used to confirm Science Motivation and Self-Efficacy Model established with the two scales. The model and coefficients obtained according to the results of CFA are given in path diagram in **Figure 1**.



**Figure 1.** Science Motivation and Self-Efficacy Model path diagram.



In addition to the values of the model above, goodness of fit statistics of the model ( $\chi^2 = 401.661$ ,  $\chi^2/df = 7.578$ ,  $RMSEA = .038$ ,  $SRMR = .034$ ,  $TLI = .995$ ,  $CFI = .996$ ) were between acceptable intervals. Researchers can carry out measurement invariance tests through different methods according to the type of the scale and variables, whether the data set has a normal distribution or not, and the size of the sample. Categorical and ordinal variables are used in this study. Moreover, univariate normality assumption is ensured while multicollinearity and homoscedasticity assumptions are not ensured thoroughly. Because of the aforementioned reasons, the method of ULS estimation, which is reported to give good results for MGCFA under the roof of CFA and SEM, was preferred in this study (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009; Kořar & Yılmaz Kořar, 2015). The analysis was carried out with the open-source statistical software called R ‘lavaan’ (Rossee, 2012), which gives the opportunity to make estimations with the method of ULS.

Measurement invariance in MGCFA was examined by means of testing four nested hierarchical model or hypothesis. These four hierarchical models are respectively listed as ‘configural invariance’, ‘weak invariance’, ‘strong invariance’ and ‘strict invariance’ (Byrne, Shavelson & Muthen, 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). While model-data fit was being examined at these stages, the values of  $\chi^2$ ,  $\chi^2/df$ ,  $RMSEA$  ( $RMSEA < .08$ ),  $SRMR$  ( $SRMR < .08$ ),  $TLI$  ( $TLI > .95$ ),  $CFI$  ( $CFI > .95$ ),  $\Delta CFI$  were taken into consideration (Browne & Cudeck, 1993; Hu & Bentler, 1999; Kaplan, 2000; Schermelleh-Engel, Moosbrugger & Müller, 2003; Schumacker & Lomax, 1996). When the fit was thought to be at an adequate level, the next step was started. First of all, the change in Chi-squared difference between two nested models ( $\Delta\chi^2$ ) was used to find out whether comparing the two models would be significant or not, as is suggested by Hirschfeld and von Brachel (2014) in their study. However,  $\Delta\chi^2$  is also a function of sample size and its usefulness has been discussed in many studies. Chi-squared difference test rejects the null hypothesis with too much power as the sample size increases. Cheung and Rensvold (2002) warned researchers that  $\Delta\chi^2$  has less value in making practical decisions about measurement invariance. One of the alternatives to  $\Delta\chi^2$  which was suggested by Cheung and Rensvold (2002) was the change in the CFI value ( $\Delta CFI$ ). Cheung and Rensvold (2002) stated the appropriate cut-offs for change in fit indices to determine. Wishing to extend Cheung and Rensvold’s research, Wu, Li and Zumbo (2007) provided extensive research about the practice of using the change in fit statistics to test for measurement invariance. Based on the studies mentioned above,  $\Delta CFI$  was conducted to make final decisions about which stage of measurement invariance model holds.  $\Delta CFI$  was examined among the two models that were more restricted when compared to each other.  $\Delta CFI$  values smaller than or equal to -0.01; indicate invariance is not satisfied (Cheung & Rensvold, 2002).

### 3. FINDINGS

Measurement invariance was examined through MGCFA under the roof of SEM at this stage of the study. Before examining measurement invariance by gender and statistical regions, CFA was carried out for these variables separately and the model-data fit was investigated. According to the results of CFA, the goodness of fit statistics were acceptable for gender groups whereas they were outside of the acceptable range in West Marmara ( $RMSEA = 0.00$ ,  $TLI = 1.001$ ,  $\chi^2 = 49.153$ ), East Marmara ( $RMSEA = 0.00$ ,  $TLI = 1.003$ ,  $\chi^2 = 38.173$ ), Middle Anatolia ( $RMSEA = 0.00$ ,  $TLI = 1.001$ ,  $\chi^2 = 39.439$ ), West Black Sea ( $RMSEA = 0.00$ ,  $TLI = 1.006$ ,  $\chi^2 = 32.644$ ), East Black Sea ( $RMSEA = 0.00$ ,  $TLI = 1.007$ ,  $\chi^2 = 37.088$ ), Northeast Anatolia ( $RMSEA = 0.00$ ,  $TLI = 1.012$ ,  $\chi^2 = 29.361$ ), and so these regions were excluded from the scope of the study.

Measurement invariance was tested in accordance with the hierarchical sort order of configural invariance, weak invariance, strong invariance and strict invariance. Results of the two change detection tests and the value of  $\Delta CFI$  were considered among the two invariance models. The examinations were kept going on until the stage where the study model provided the invariance in the related group.

### 3.1. Measurement Invariance of the Model by Gender

Within the scope of this model, a model was created with the two scales belonging to the PISA 2015 cycle and the measurement invariance of this model was examined by gender. At the end of the examination, which was done via the method of MGCFA, the Model of Science Motivation and Self-Efficacy met the conditions of all stages of invariance in PISA 2015 Turkey sample. According to this result, it can be stated that the item-factor structure in this study had an equal distribution among males and females. Also, factor loadings, variances, covariances and error variances are found out to be equal by gender. It can be concluded that the study model provided measurement invariance. According to Hirschfeld and von Brachel (2014), the change in the value of  $\chi^2$  between the models should be tested. The results of the test regarding the change in the value of  $\chi^2$  between the four hierarchical models are given in Table 5.

**Table 5.** Test of change in  $\chi^2$  between the four hierarchical models.

	$\chi^2$	<i>df</i>	$\Delta\chi^2$	$\Delta df$	<i>p</i>
Configural	426.641	106	-	-	
Weak	452.303	116	25.66	10	0.004
Strong	625.290	138	172.99	22	<.01
Strict	1114.154	140	488.86	2	<.01

Moreover, the change in  $\chi^2$  value of the models created in four different stages was examined. As is seen in Table 5, the change in Chi-square was found to be 25.66 and this change is statistically significant at the level of  $p < .01$ . Although the goodness of fit statistics were between acceptable intervals, the change in Chi-square was found to be significant. However, it is known that Chi-squared tests are highly sensitive to sample size. For this reason, the study was continued with the analyses of hierarchical models. The stages that were tested during the examination of measurement invariance by gender and the data belonging to the goodness of fit statistics are given in Table 6.

**Table 6.** Measurement invariance of the model by gender.

	$\chi^2$	<i>df</i>	$\chi^2/df$	<i>RMSEA</i>	<i>SRMR</i>	<i>TLI</i>	<i>CFI</i>	$\Delta CFI$
Configural	426.641	106	4.02	0.036	0.035	0.995	0.996	-
Weak	452.303	116	3.90	0.036	0.036	0.995	0.996	0.000
Strong	625.294	138	4.53	0.039	0.036	0.994	0.994	-0.002
Strict	1114.154	140	7.98	0.055	0.036	0.988	0.988	-0.006

When the goodness of fit statistics and  $\Delta CFI$  ( $\Delta CFI > -.01$ ) values given in Table 6 are considered ( $RMSEA < .08, SRMR < .08, TLI > .95, CFI > .95$ ), it can be concluded that measurement invariance is ensured between genders.

### 3.1. Measurement Invariance of the Model by Statistical Regions

The model created with the two sub-scales was examined among statistical regional groups. At the end of the examination that was done through the method of MGCFA, the Model of Science Motivation and Self-Efficacy met the conditions of all stages of invariance in PISA 2015 Turkey sample. Before having the data of the goodness of fit statistics of the models, the change

in the value of  $\chi^2$  between the models were tested. The results of the test regarding the change in the value of  $\chi^2$  between the four hierarchical models are given in [Table 7](#).

**Table 7.** Test of change in  $\chi^2$  among the four hierarchical models.

	$\chi^2$	df	$\Delta\chi^2$	$\Delta df$	p
Configural	446.719	318	-	-	-
Weak	501.565	368	54.846	50	0.296
Strong	698.656	478	197.091	110	<.01
Strict	997.433	488	298.77	10	<.01

The change in  $\chi^2$  value between the models which were obtained at the end of each stage and which were more restricted when compared to each other was examined. p value that was obtained at the stage of strong invariance is statistically significant at the level of 0.01. When [Table 7](#) is examined, it is clear that the change between the Chi-square values of weak and strong invariance models is found to be 197.091 and this change is statistically significant at the level of  $p < .01$ . Although the goodness of fit statistics were between acceptable intervals, the change in Chi-square was found to be significant. However, it is known that Chi-squared tests are highly sensitive to sample size. For this reason, the study was continued with the analyses of hierarchical models, as Wu, Li and Zumbo (2007) suggested. The stages that were tested during the examination of measurement invariance by statistical regions and the data belonging to the goodness of fit statistics are given in [Table 8](#).

**Table 8.** Measurement invariance of the model by statistical regions.

	$\chi^2$	df	$\chi^2/df$	RMSEA	SRMR	TLI	CFI	$\Delta CFI$
Configural	446.719	318	1.40	0.028	0.042	0.997	0.998	-
Weak	501.565	368	1.63	0.026	0.044	0.997	0.998	0.000
Strong	698.656	478	1.46	0.030	0.045	0.997	0.996	-0.002
Strict	997.433	488	2.04	0.044	0.046	0.996	0.991	-0.005

When the goodness of fit statistics and  $\Delta CFI$  ( $\Delta CFI > -.01$ ) values given in [Table 8](#) are considered ( $RMSEA < .08, SRMR < .08, TLI > .95, CFI > .95$ ), it can be concluded that measurement invariance is ensured among the statistical regions.

#### 4. DISCUSSION and CONCLUSION

A model was created in this study with the two sub-scales included in the PISA 2015 Student Questionnaire and the measurement invariance of this model was examined among different groups. It was found out at the end of the study that the model could provide measurement invariance by gender and statistical regions. The Model of Science Motivation and Self-Efficacy shows that valid comparisons can be made among the scores of male and female students as well as students in different regions of Turkey. The study results comply with the results of the study carried out by Kıbrıslıoğlu Uysal and Akın Arıkan (2018), who examined the measurement invariance of the Science Self-Efficacy Scale used in PISA 2006 and 2015 by gender. The researchers reported at the end of the study that the scale that was used in the two PISA cycles met the conditions of all the stages of measurement invariance by gender. It seems necessary to be more careful while preparing assessment tools to be used in affective domains in science and to carry out measurement invariance examinations for these tools in different groups. Moreover, it would be helpful for researchers to make comparisons by gender and statistical regions with the scores obtained through the assessments in affective domains in science. On contrary to this, Uyar and Kaya Uyanık (2019) used the affective qualities about science and gender as variables in their study that was conducted with PISA 2015 Turkey

sample. The results of their analyses showed that their model provides only configural and weak invariance between genders in Turkey sample. In addition, Uyar and Doğan (2014) used same variables to investigate measurement invariance of a model on learning strategies in ‘Learning by strategies’ part of PISA 2009 Student Questionnaire. In their study, it is reported that while the model only provided configural and weak invariance stages in the groups of gender and school types, it provided all measurement invariance stages among regions of Turkey.

In this study, measurement invariance by gender and statistical regions of the created model was examined. As a result, it is found that strict measurement invariance was provided by either of the variables. Measurement invariance was examined through MGCFA under the roof of SEM in this research. It is possible to have different results when different methods are used in measurement invariance examinations (Yandı, Köse & Uysal, 2017). It is recommended that the research model should be tested with different methods. Moreover, four hierarchical nested models were tested while examining measurement invariance with MGCFA. There are different notions in the literature about the comparison of the nested models and the rejection of the null hypothesis. At those stages,  $\Delta\chi^2$  and  $\Delta CFI$  were used in addition to the goodness of fit statistics. It is possible to have different results, when  $\Delta\chi^2$  and  $\Delta CFI$  values are used together. In this study,  $\Delta CFI$  values were used to make final decisions about which stage of measurement invariance model holds.  $\Delta CFI$  values were examined among the two models that were more restricted when compared to each other, as Wu, Li and Zumbo (2007) suggested. More research results are needed about which criteria can be accepted while making comparisons in the investigation of measurement invariance. It is thought that simulation studies could provide this contribution to the field.

Although there are a number of studies about the cognitive domain in literature, the researchers that carry out in the affective domain state that they do research in a more virgin field (Boyd, Dooley & Felton, 2006). There is a need for study results that will contribute to developing scales in this field as well as analyzing the differences among groups in the country. It is believed that this study will contribute to the literature in this field with its results. In this study, measurement invariance was conducted only for gender and statistical regions. Future research on the invariance of the construct across different demographic groups would be concerning.

### Acknowledgements

This paper was produced from first author’s master thesis.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

### ORCID

Metehan Güngör  <https://orcid.org/0000-0003-4409-2229>

Kübra Atalay Kabasakal  <https://orcid.org/0000-0002-3580-5568>

### 5. REFERENCES

- Ağaç, G., & Masal, E. (2015). An investigation of the relation between 8th grade students’ beliefs, abstract thought and achievement; The case of mathematics. *International Online Journal of Educational Sciences*, 7(1), 134-144.
- Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Van Yuzuncu Yil University Journal of Education*, 6(1), 18-47.

- Alivernini, F. (2011). Measurement invariance of a reading literacy scale in the Italian context: A psychometric analysis. *Procedia Social and Behavioral Sciences*, 15, 436-441.
- Anderson, C. J., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 3, 411-423.
- Asil, M., & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği [Cross-cultural equivalence of the PISA student questionnaire]. *Education and Science*, 37(166), 236-249.
- Ayvallı, M., & Biçak, B. (2018). An investigation into the measurement invariance of PISA 2012 mathematical literacy test. *European Journal of Education Studies*, 4 (11), 39-58.
- Bakan Kalaycıoğlu, D. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory & Practice*, 15(5), 1-11.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.
- Başusta, B. N. (2010). Ölçme eşdeğerliği [Measurement equivalence]. *Journal of Measurement and Evaluation in Education and Psychology*, 1(2), 58-64.
- Başusta, B. N., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA anketi örneği [Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire]. *Hacettepe University Journal of Education*, 30(4), 80-90.
- Batyra, A. (2017). Gender gaps in student achievement in Turkey: Evidence from Trends in International Mathematics and Science Study (TIMSS) 2015. *Education Reform Initiative & Aydın Doğan Foundation*.
- Bıkmaz, H. F. (2002). Fen öğretiminde öz-yeterlik inancı ölçeği [Self-efficacy belief instrument in science teaching]. *Educational Sciences & Practice*, 1(2), 197-210.
- Bircan, H., & Sungur, S. (2016). The role of motivation and cognitive engagement in science achievement. *Science Education International*, 27(4), 509-529.
- Boyd, L. B., Dooley, E. K., & Felton, S. (2006). Measuring learning in the affective domain using reflective writing about a virtual international agricultural experience. *Journal of Agricultural Education*, 47(3), 24-32.
- Britner, S. L. & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering*, 7(4), 271-285.
- Browne, M. W., & Cudeck, R. (1993). *Alternative ways of assessing model fit, testing structural equation models*, K. A. Bollen & J. S. Long (Eds.), Newbury Park, CA: Sage.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.
- Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-cultural Psychology*, 34(2), 155-175.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2016). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Ankara: Pegem.
- Dilts, R. (1998). Motivation. Retrieved August 17, 2019 from <http://nlpu.com/Articles/artic17.htm>
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5, 23-35.
- Ertürk, Z., & Erdiñç Akan, O. (2018). TIMSS 2015 matematik başarısı ile ilgili bazı değişkenlerin cinsiyete göre ölçme değişmezliğinin incelenmesi [The investigation of

- measurement invariance of the variables related to TIMSS 2015 mathematics achievement in terms of gender]. *Journal of Theoretical Educational Science, UBEK-2018*, 204-226.
- Forero, G. C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16, 625-641.
- Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, 25(4), 280-296.
- 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(11), 78-94.
- Gujarati, D. N. (1995). *Basic econometrics (3rd Ed.)*. New York, NY: Mc-Graw Hill.
- Gülleroğlu, D. H. (2017). PISA 2012 matematik uygulamasına katılan Türk öğrencilerin duyuşsal özelliklerinin cinsiyete göre ölçme değışmezliğinin 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*, 37(1), 151-175.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.
- Hirschfeld, G., & Brachel, v. R. (2014). Multiple-group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, 19(7), 1-12.
- Hu, L., & Bentler, M. P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- İlhan, K. (2015). *Eğitimde pozitif psikoloji uygulamaları*. B. Ergüner Tekinalp & Ş. Işık (Ed.). Ankara: Pegem Akademi Yayıncılık.
- İmrol, F. (2017). Investigation of measurement invariance of motivation and self-belief constructs towards mathematics in PISA 2012 Turkey sample (Master thesis). Retrieved August 14, 2019 from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 user's reference guide*. Chicago: Science Software International.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage.
- Karakoç Alatlı, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 fourth grade mathematics test in terms of cross-cultural measurement invariance. *Euroasian Journal of Educational Research*, 66, 389-406.
- Kıbrıslıoğlu, N. (2015). The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey – China (Shanghai) - Indonesia (Master thesis). Retrieved August 14, 2019 from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- 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.
- Kline, R. B. (2005). *Principles and practices of structural equation modeling (2nd Ed.)*. New York: Guilford Press.
- Kline, R. B. (2011). *Principles and practices of structural equation modeling (3rd Ed.)*. New York: Guilford Press.

- Koğar, H., & Yılmaz Koğar, E. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 351-364.
- Koh, H. K., & Zumbo, D. B. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods*, 7(2), 471-477.
- Larson, L. M., Stephen, A., Bonitz, V. S., & Wu, T.-F. (2014). Predicting science achievement in India: role of gender, selfefficacy, interests, and effort. *Journal of Career Assessment*, 22(1), 89-101.
- Mellenberg, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research: Applications of Item Response Theory*, 13(2), 123-144.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.
- Meredith, W., & Millsap, E. R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57(2), 289-311.
- Ministry of National Education [MNE]. (2016). PISA 2015 national report. Ankara.
- Nagengast, B., & Marsh, H. (2014). *Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006*. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp 317-345). UK: Taylor & Francis.
- OECD [Organisation for Economic Co-operation and Development]. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. PISA, OECD, Paris: OECD Publishing.
- OECD [Organisation for Economic Co-operation and Development]. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy and Collaborative Problem Solving, revised edition*. PISA, Paris: OECD Publishing.
- Oliden, E. P., & Lizaso, M. J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema*, 25(3), 390-395.
- Ölçüoğlu, R. & Çetin, S. (2016). TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi [The investigation of the variables that affecting eight grade students' TIMSS 2011 math achievement according to regions]. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 202-220.
- Polat, E., & Madra, A. (2018). PISA 2015 ve TIMSS 2015 ışığında Türkiye’de cinsiyete dayalı başarı farkı. *Education Reform Initiative & Aydın Doğan Foundation*, 1-18.
- Raykov, T., & Marcoulides, A. G. (2006). *A first course in structural equation modeling*. Lawrence Erlbaum Associates, Inc.
- Reise, P. S., Widaman, F. K. & Pugh, H. R. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552-566.
- R Development Core Team (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved August 14, 2019 from <http://www.R-project.org/>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36.
- Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Front. Psychol.*, 7(110), 1-16.

- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.
- Schumacker, E. R., & Lomax, G. R. (1996). *A beginner's guide to structural equation modeling*. Manwah, NJ: Lawrence Erlbaum Associates.
- Schunk, D. H., Meece, J. L. & Pintrich, P. R. (2014). *Motivation in education: Theory, research and applications*. New Jersey: Pearson Education, Inc.
- Schunk, D. H., & Pajares, F. (2009). *Self-efficacy theory*. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of Motivation at School* (pp. 35-53). New York, NY: Routledge.
- Sireci, S. G., Patsula, L. N., & Hambleton, R. K. (2005). *Statistical Methods for Identifying Flaws in the Test Adaptation Process*. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. New Jersey, London: Lawrence Erlbaum Associates, Publishers.
- Somer, O., Korkmaz, M., Dural, S., & Can, S. (2009). Ölçme eşdeğerliğinin yapısal eşitlik modellemesi ve madde cevap kuramı kapsamında incelenmesi [Detection of measurement equivalence by structural equation modeling and item response theory]. *Turkish Journal of Psychology*, 24(64), 61-75.
- Steenkamp, J.B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.
- Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.
- Taris, W. T., Bok, A. I., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: a general approach. *The Journal of Psychology*, 132(3), 301-316.
- Uyar, Ş., & Doğan, N. (2014). 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]. *International Journal of Turkish Education Sciences*, 3, 30-43.
- Uyar, Ş., & Kaya Uyanık, G. (2019). Fen bilimlerine yönelik öğrenme modelinin ölçme değişmezliğinin incelenmesi: PISA 2015 örneği [Investigation measurement invariance of learning model towards science: PISA 2015 sample]. *Kastamonu Education Journal*, 27(2), 297-507.
- Uzun, N. B., Gelbal, S., & Öğretmen, T. (2010). TIMSS-R fen başarısı ve duyuşsal özellikler arasındaki ilişkinin modellenmesi ve modelin cinsiyetler bakımından karşılaştırılması [Modeling the relationship between TIMSS-R science achievement and affective characteristics and comparing the model according to gender]. *Kastamonu Education Journal*, 18(2), 531-544.
- Yandı, A., Köse, A. İ., & Uysal, Ö. (2017). Examining measurement invariance with different methods: Example of PISA 2012. *Mersin University Journal of the Faculty of Education*, 13(1), 243-253.
- Vandenberg, J. R., & Lance, E. C. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68-81.
- 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 & Evaluation*, 12(3), 1-26.
- Zedlin, A. L., Britner, S. L., & Pajares, F. (2007). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, 45(9), 1036-1058.