International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 325-338

DOI: 10.21449/ijate.379508

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

Measurement Invariance of Science Self-Efficacy Scale in PISA

Nermin Kıbrıslıoğlu Uysal [1,*], **Çiğdem Akın Arıkan** [1]

[1]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation, Beytepe, Ankara - Turkey

Abstract: The aim of this study is to find out whether the science self-efficacy scale in PISA 2006 and PISA 2015 applications ensure measurement invariance. Sample of the study consists of 4791 students in PISA 2006 and 5071 students in PISA 2015 implementation. Multi-group Confirmatory Factor Analysis (MGCFI) was performed to determine invariance of the science self-efficacy scale across year and gender. Invariance stages were examined by means of the comparison of fit indexes. The results of the study indicated that the science self-efficacy scale satisfied all stages of invariance by gender both in 2006 and 2015. Structural and metric invariance was provided for both gender across years and total group across years.

ARTICLE HISTORY

Received: 16 January 2018
Revised: 11 April 2018
Accepted: 14 April 2018

KEYWORDS

PISA, Self-efficacy, Gender, Measurement invariance

1. INTRODUCTION

In recent years, interaction and competition among countries has been growing with development in technology and ease of communication. This situation is accompanied by comparisons made among countries in many areas (OECDa). Therefore, international organizations carry out studies in different areas. These works not only allow countries to assess themselves in the international platform but also constitute feedback regarding their policies and education levels.

A particular work done by the OECD in the field of education is the PISA (Program for International Student Assessment). PISA has been applied every three years since 2000, and its main goal is to measure whether 15-year old students, who are expected to complete mandatory education use the information obtained from their education life in real life situations independent from countries' educational curricula (OECDb). Turkey has participated in PISA since 2003. In Turkey, while PISA was given as a pencil and paper exam until 2015, it was turned into computer-based application in 2015 (OECD, 2015).

PISA evaluations include reading, mathematics and science test in the cognitive domain and each year one of these areas is taken as the main focus. Besides these tests, within the scope of PISA applications, surveys are given to students, parents and school administrators

CONTACT: Nermin Kıbrıslıoğlu Uysal ✉ nkibrislioglu@hacettepe.edu.tr Hacettepe Üniversitesi, Beytepe Yerleşkesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı 06800, Beytepe, Ankara

ISSN-e: 2148-7456 /© IJATE 2018

in order to learn more broadly about students' backgrounds and learning experiences and school system and learning environments. Such surveys also focus on the core area covered in each cycle (OECDb). Thus, the same basic area is measured every 9 years. Apart from that, in order to reveal the trend by years and associate the findings with previous tests in PISA applications; common items are used in the cognitive domain and common scales in the affective domain. Thus, students are compared by years in both cognitive and affective domains (OECDb).

PISA 2006 and 2015 implementations focus on science. Though revised partially, the students' surveys applied in 2006 and 2015 contain similar scales. Therefore, it is possible to make a comparison on affective characteristics of students across years. In addition, the literature provides abundant samples of comparison of affective characteristics by gender. However, in order to derive the correct results from the comparisons made, the examinees who are equivalent in terms of the trait measured must get the same score from the tests or scales (Schimit and Kuljanin, 2008). To put it differently, the same trait must be associated identically with the group of variables observed the same in all groups (Borsboom, 2006). In other words, the scores obtained from the scales can be used for comparison across different groups provided that the scales ensure the measurement invariance between the groups concerned. In fact, measurement invariance is an assumption that must be checked before comparisons are made between groups because the traits measured where measurement invariance is not met, may not be identical across the groups measured (Vandenberg & Lance, 2000). Nonetheless, measurement invariance is rarely tested in studies. This makes the validity of the results obtained questionable because comparisons are made without having information about construct validity of scales and equality of the validity across groups (Gregorich, 2006).

1.1. Measurement Invariance

In general terms, measurement invariance refers to examining whether the scores measuring a particular construct have the same meaning under different circumstances. Different conditions could include different populations, different measurement times or different methods of application (such as paper-pencil and computer-based). Consistency of the construct across years is referred to as longitudinal measurement invariance and deals with whether factor structure varies in longitudinal pattern by years. Invariance between populations is related to structural bias and investigates whether the measured trait is the same or not across groups (Kline, 2011).

A widely used method for measuring invariance is the multi-group confirmatory factor analysis (MGCFI) method (Widaman and Rice, 1997; Vandenberg and Lance, 2000; Kline, 2011). In the MGCFI method, different stages of measurement invariance can be tested for different purposes. Those stages can be listed as invariance of covariance matrice, configural invariance, metric invariance, scalar invariance, strict factorial invariance, invariance of factor covariance and invariance of factor averages (Vandenberg & Lance, 2000). As a matter of fact, the comparability of observed scores between groups can be provided with configural, metric, scalar and strict factorial invariance (Widaman & Rice, 1997). In this study, measurement invariance is investigated in relation with these four type of invariance. The stages of measurement invariance are summarized in Table 1**.**

As seen in Table 1, the first stage of measurement invariance is configural invariance. Configural invariance only requires the identical measurement pattern across groups. If configural invariance is not provided, measurement invariance will not be ensured at any stage (Kline, 2011). Secondly, metric invariance requires identical factor loadings across groups as well as configural invariance. Metric invariance is also called weak factorial invariance. Once metric invariance is ensured, it can be argued that the covariance differences in the variable measured across groups arise from the common factors; leaving the root of observed score differences between groups unexplored (Millsap & Olivera-Aguilar, 2012). On the other hand, when metric invariance is not ensured, it could be argued that the factors do not have the same meaning across groups (Gregorich, 2006). The following stage, scalar invariance is a strong level of invariance and requires equality of factor variance and covariances between groups as well as metric invariance. When scalar invariance is ensured, comparison of differences between averages of the groups yields significant outcomes (Millsap and Olivera-Aguilar, 2012). Finally, strict factorial invariance requires equality of item residual variances between groups in addition to scalar invariance. Provision of strict factorial invariance leads the way for comparing not only observed variable averages but also factor variance and covariances between groups (Gregorich, 2006). However, as variance of the latent variable increases, item residual variance also increases, strict factorial invariance is often not achieved in practice. The stages of measurement invariance are hierarchical. Therefore, the stages are evaluated respectively, and if invariance is not provided at any stage, there is no need to examine the following stage.

Table 1. Measurement Invariance Stages

Degree of Invariance	Condition of Invariance	Group Comparison
Configural invariance	Item/Factor groups	---
Metric invariance	Item/Factor groups and factor loads	Factor variance and covariances
Scalar Invariance	Item/Factor groups, factor loads and item constants	Factor variance and covariances, factor and observed variable averages
Strict factorial invariance	Item/Factor groups, factor loads, item constants, and item residual variances	Factor variance and covariances, factor and observed variable averages, observed variance and covariances

1.2. Self-efficacy

According to Bandura (1982), self-efficacy is the self-judgment about how well an individual can do a behavior. Self-efficacy perception affects behaviour and performance, as well as beliefs of individuals. So even if individuals have an idea about the result of a behavior, they tend to avoid conducting that behaviour as long as they have a low level of self-efficacy related to that particular behaviour (Bandura, 1977).

Self-efficacy is not a personal trait by nature; rather, it focuses on performance capabilities targeting specific objects (Zimmerman, 2000). In this case, science self-efficacy can be defined as the extent at which students believe in their own abilities to succeed in science-related tasks. Self-efficacy has an impact on future-oriented behaviours of individuals. In other words, before individuals perform any behavior, they evaluate their self-efficacy towards that behavior. In this regard, students' self-efficacy perception in a particular subject area affects their desire to underatake activities related to that field, the efforts they would show for these activities and continuity of the efforts, and thus their performance in that area (Zimmerman, 2000).

Students' science self-efficacy affects their desire to undertake science related tasks, science achivement as well as their future preferences (Lent, Brown, & Larkin, 1986; Post, Stewart & Smith, 1991; Andrew, 1998; Scott & Mallinckrodt, 2005; Zedlin, Britner & Pajares, 2007). For example, Scott ve Mallinckrodt (2005) examined female high school students' science self-efficacy and their career preferences related to science. They reported that between strudents who prefereed science related major and the ones do not, differ significantly with respect to their science self-efficacy. Moreover, there are plenty of studies related to comparison of male and female students' self-efficacy in the literature (Post, Stewart & Smith, 1991; Britner & Pajares, 2001; Zedlin, Britner & Pajares, 2007:) Britner and Pajares (2001) investigated possible gender differences on high school students' science self-efficacy and motivation. They reported that girls have stronger science self-efficacy beliefs and higher grades while boys have stonger performance-approach goals. Zedlin, Britner and Pajares (2007) examined the self-efficacy beliefs of men and women who selected career in science and mathematics majors. The results of the study indicated that women and men have different sourses of self-efficacy beliefs.

Hence, science self-efficacy is an important construct and it is studied in the literature frequently. Moreover, comparisons of science self-efficacy between gender groups is also very common. In order to provide correct implications from these comprasions it is very important to test the invariance of these scales (Vandenberg & Lance, 2000).

In order to determine scientific literacy in PISA applications, not only achievement tests but also surveys on students' affective traits related to academic achievement are applied. In both 2006 and 2015 tests, students' science self-efficacy was measured in the scope of the affective domain surveys. The scale items were same in 2006 and 2015. In relation with science self-efficacy, a number of tasks were listed in the students' survey and students were asked how easy they find it to do these tasks on their own (MEB, 2010). The scale items are shown in Table 2.

Table 2. PISA 2006-2015 science field self-efficacy scale items

1. Recognizing the question underlying the newspaper article on a health problem
2. Explaining why earthquakes take place more often in some areas
3. Explaining the role of antibiotics in treatment of diseases
4. Identifying the problem regarding proper collection and treatment of wastes
5. Predicting how the changes in the environment could affect survival of certain living species
6. Interpreting the scientific information on labels on foodstuffs
7. Discussing how new evidence can change the understanding that there is life on the Mars
8. Deciding which of the two views about acid rains is better

*Taken from 2006 and 2015 PISA National Reports.

1.3. Aim of the Study

The aim of this study is to find out whether or not the science self-efficacy scale, which was given as a common scale in 2006 and 2015, satisfies measurement invariance across years and gender groups in Turkey sample.

In the literature, there are many studies related to measurement invariance for different groups on scales used in international tests such as PISA and TIMSS (Ercikan and Koh, 2005; Marsh et al., 2006; Wu, Lin & Zumbo, 2007; Lee, 2009; Akyıldız, 2009; Uzun & Öğretmen, 2010; Güzeller, 2011; Asil & Gelbal, 2012; Uyar & Doğan, 2014; Başusta and Gelbal, 2015; Bulut, Palma, Rodrigez and Stanke, 2015; Kıbrıslıoğlu, 2015; Karakoç and Alatlı, 2016; Ölçüoğlu & Çetin, 2016; Gülleroğlu, 2017). Ercikan and Koh (2005) examined invariance of English and French forms in TIMSS 1999 implementation. They analysed the invariance with both MGCFA and differential item functioning (DIF) analysis. They reported that both mathematics and science forms did not ensure measurement invariance. Lee (2009) examined whether math self-concept, math self-efficacy, and math anxiety scales in PISA 2003 implementation provide one consistent factor structure between 41 countires. For this purpose, he conducted exploratory, confirmatory and multi group confirmatory factor analysis. The results of the study indicated that structure of these constructs differ between countries. Bulut, Palma, Rodrigez and Stanke (2015) investigated measurement invariance of support and positive identity scales among White and Latin American students across years. The results of the study indicated that subgroup-year interaction has a significant effect on parameter shift. Hence, the invariance of scale parameters between two different groups of students differs across years did not provided.

Uyar and Doğan (2014) investigated measurement invariance of the model for learning strategies in the PISA 2009 students' survey across gender, school type and statistical region group. It was found out that only configural invariance and metric invariance were met in gender and school type groups, while all of the invariance conditions were fulfilled in relation with regions. In another study, Uzun and Öğretmen (2010) investigated whether variables affecting students' success in science such as self-efficacy, attitude, significance and in-class student activities satisfy measurement invariance in the Turkish participants in 1999 TIMMS-R. They found out that self-efficacy, significance and in-class student activities satisfy metric invariance; while attitude satisfies scalar invariance between gender groups. Kıbrıslıoğlu (2015) investigated invariance of the items in mathematics subscale of PISA 2012 was investigated across gender and cultures. It was found out that intercultural invariance meets invariance only in configural level while gender groups meets strict factorial invariance. Furthermore, Başusta and Gelbal (2015) examined measurement invariance of the science and technology items in the PISA 2009 students' survey against gender. They reported that scalar invariance ensured between gender groups. Also, Gülleroğlu (2017) investigated measurement invariance of affective traits such as interest, anxiety, self-efficacy and sense of self regarding mathematics against gender in the PISA 2012 implementation. It was noted that mathematics self-efficacy scale does not satisfy configural invariance. On the other hand, sense of self-regarding mathematics scale ensured configural invariance and anxiety and interest towards mathematics satisfy scalar invariance.

Different from the literature, present study investigates not only measurement invariance across years and genders but also whether or not the scale items satisfy invariance for each gender between the years 2006 and 2015. Due to the non-longitudinal nature of PISA data, the analyses targeted measurement invariance in different groups across years. In the PISA applications, population defined as the representative population and the sample is selected at random. Present study was carried out assuming that both applications consisted of samples with similar individuals. Therefore, the study is expected to demonstrate whether or not the construct varied in gender subgroups between 2015 and 2006. Moreover, investigation of bias in subgroups allow for unearthing the probable bias that can not be revealed as a result of bias analysis for the whole group but affects subgroups (Huggings-Manley, 2016). So, the study is considered significant as it attempts to additionally reveal the change between genders over the years.

Hence, this study was intended to find out whether science self-efficacy scale in PISA 2006 and 2015, which were given as common tests, satisfy measurement invariance depending on year and gender, and also invariance against year in gender subgroups. Answer was sought for the following questions in the study:

(1) Does science self-efficacy scale satisfy measurement invariance between the years 2006 and 2015?

(2) Does science self-efficacy scale satisfy measurement invariance between gender subgroups?

2. METHOD

2.1. Research Method

This study is a descriptive study as it aims identifying whether the science self-efficacy scale in the students' survey in the PISA 2006 and PISA 2015 implementations is invariant by years and gender.

2.2. Population and Sample

A total of 4942 students from 160 schools participated in the PISA 2006 application held in Turkey. The participants were selected from 7 geographic regions, 51 provinces in a random manner by two-step stratification of regions and schools. As for PISA 2015, participants were selected by means of two-stage random sampling method. At step one; schools and students were identified by means of stratified random sampling with respect to the strata of Statistical Region Units Classification (SRUC) Level 1, education type, school type, the place of schools and administrative form of schools. PISA 2015 was administered to 5895 students from 187 schools in 61 provinces to represent 12 different regions according to the SRUC Level 1 (MEB, 2017).

2.3. Data Analysis

Measurement invariance between the groups was examined by means of MGCFI method. Before analysis, the assumptions of missing data, extreme values, multivariate normality, and multicollinearity were tested (Çokluk, Şekercioğlu & Büyüköztürk, 2012). The assumptions are elaborated below.

For missing data, first of all, the examinees who responded to none of items in the PISA 2006 and PISA 2015 applications were removed from the data. Then, missing data analysis was performed for both data sets and missing data rates were examined. The analysis yielded missing data rate below 5% distributed randomly. Kline (2011) stated that in the case of large samples, missing data rate below 5% with random distribution, such data could be omitted. For this reason, the missing data were removed from both data sets under listwise deletion condition. Analyses were performed for 4814 and 5235 respondents for the PISA 2006 and PISA 2015 implementations, respectively.

Secondly, univariate and multivariate outliers were examined. One way to determine univariate extreme values is to convert variables into standard variables. In large samples (n> 100), z-scores outside the range of -3 to +3 are regarded extreme values (Tabachnick & Fidell, 2007). On the other hand, multivariate extreme values can be computed from the Mahalanobis distance. The Mahalanobis distance exhibits the chi-square distribution and degree of freedom is equal to the number of variables in the data set. The values smaller than the chi square value at 0.001 significance level were identified as outliers (Tabachnick and Fidell, 2007). For univariate extreme values, z values were examined indicating no case outside the specified range. Mahalanobis distances were then investigated for multivariate extreme values. As a result, 23 respondents were removed from analysis as extreme values for PISA 2006 and 164 respondents for PISA 2016, respectively. Frequency table for the remaining respondents are given in Table 3.

Table 3. Gender-Related Frequency

Gender	2006		2015	
	f	%	*f*	%
Female	2229	46.5	2559	50.5
Male	2562	51.5	2512	49.5
Total	4791	100	5071	100

It can be seen in Table 3 that 4791 participants in year 2006 included 2229 females and 2562 males; whereas 5071 participants in 2015 consisted of 2559 females and 2512 males. The distribution of gender groups by years is balanced.

Multivariate normality is achieved provided that univariate normality is provided and linearity and residuals are covariant (Kline, 2011). For univariate normality assumption, skewness and kurtosis coefficients of the variables were examined. It was found out that the skewness and kurtosis coefficients for PISA 2015 data are in the range of -1 and +1. As for the PISA 2006 data, only the coefficient of item 7 was found to be -1.064 the other variables falling in the specified range. The skewness and kurtosis coefficients in the specified range suggest that the variables satisfy normality assumption (Büyüköztürk, 2002). As for linearity, residual graphs were examined indicating that linearity assumption is met. For homoscedasticity, Durbin Watson values were examined with resulting values in the range of 0 - 4 fulfilling the homoscedasticity assumption (Tabachnick & Fidell, 2007). In addition, multivariate normality was checked with Bartlett sphericity Test yielding significant results for all subgroups. This shows suitability of the data for multivariate normal distribution (Çokluk et al., 2012).

Multicollinearity assumption was tested with tolerance value, conditional index and variance inflation factor values (VIF). Multicollineraity does not exist when tolerance values are greater than 0.10 for the absence of transactions, VIF values are smaller than 10, and conditional index values are (CI) smaller than 30 (Tabachnick & Fidell, 2007). In this study, tolerans values are between 0.4-0.8, VIF values are between 1.5-2.4 and CI values are between 1-12.5. Analyses showed no problem of multicollinearity in data. The investigations revealed that the data meet the required assumptions. Prior to MGCFI, a confirmatory factor analysis (CFA) was performed and goodness of fit statistics were examined to test the fit of the model. The CFA model for PISA 2006 is shown in Figure 1.

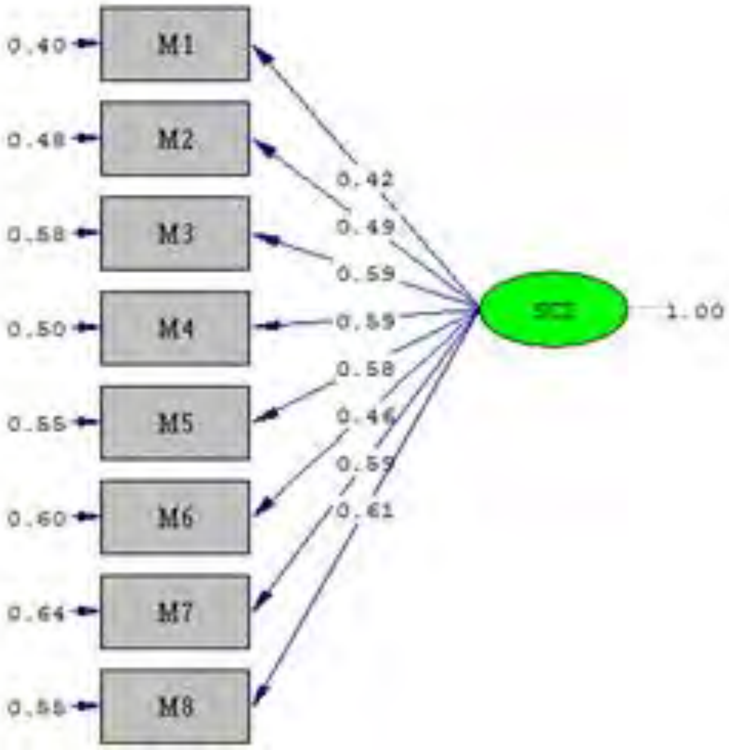

Figure 1. Configural model of the science self-efficacy scale

The values in the model above shows that factor loadings fall in the range of .40 and .64. It corresponds to compliance indices at acceptable levels except from χ2/df ratio (χ2= 427.35, df= 20, CFI=0.98, TLI= 0.97 RMSEA= 0.06, GFI=0.97). The chi-square statistic is affected by sample size and usually significant in large samples (Kline, 2011) and therefore, χ2 value does not taken as a basis for rejection or acceptance of the models (Schermelleh-Engel, Moosbrugger & Müller, 2003). So, the model was evaluated according to other indices.

MGCFA analysis was conducted in four stage. At first; configural invariance, which has free factor loads, factor correlations and error variances was tested. At the following stage, metric invariance was tested, which has free factor correlations and error variances under the condition of equal factor loads. Then, scalar invariance was tested with equal factor loads and factor correlations but free error variances. Lastly, strict factorial invariance was tested which has equal factor loads, factor correlations and error variances. At each stage, the difference values of the comparative fit index, (ΔCFI) were examined to decide whether invariance is satisfied or not. ΔCFI values smaller than or equal to -0.01; indicates invariance is achieved; otherwise it is not satisfied (Cheung & Rensvold, 2002).

3. RESULTS

This study was carried out to investigate measurement invariance of science self-efficacy scale over years as well as gender groups. The findings are reported in a way to discuss the research problems one by one.

3.1. MGCFI Results by Years

The goodness of fit indexes obtained at each invariance stage of the MGCFI are displayed in Table 4, which implies whether the science self-efficacy scale consists of eight items are equivalent between the years 2006 and 2015.

Table 4. Goodness of fit indexes by levels of invariance for 2006 and 2015

	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	1595.4	40	0.092	0.039	0.96	0.97	
Metric	1693.2	47	0.087	0.043	0.97	0.97	0
Scalar	2734.0	63	0.094	0.096	0.96	0.96	-0.01
Strict	2596.6	71	0.087	0.055	0.98	0.97	0.01

The goodness of fit indexes in Table 4 show that the RMSEA value was outside the acceptable interval at the configural invariance stage, while the other statistics of concordance fall in the acceptable range (RMSEA>.08; SRMR<.1; TLI>.95; CFI>.95). This means that the structure of the model has remained the same over the years. After providing configural invariance, metric invariance was tested.

For metric invariance, the goodness of fit indexes in Table 4 show that the RMSEA value is outside the acceptable range, but the other statistics indicate model fit. Metric invariance is ensured as ΔCFI value is in acceptable range (ΔCFI ≤ 0.01). This finding implies that the relations between the measured traits and self-efficacy dimension have remained similar across years. As metric invariance ensured, scalar invariance is tested.

The values in Table 4 show that RMSEA values are outside the acceptable range, while the rest of the statistics fall within the acceptable range. ΔCFI value indicate scalar invariance was met, (ΔCFI ≤ 0.01). Hence we concluded that sclalar invariance was ensured. This finding suggests that item factor loads and factor correlations are similar in both years. After ensuring scalar invariance, the last stage, strict factorial invariance, was implemented.

While checking the indices in Table 4 for strict invariance, the differences between CFI were obtained for scalar invariance and strict factorial invariance, respectively. The ΔCFI values reveal that strict factorial invariance is satisfied in this case.

The analysis of the invariance between 2006 and 2015 in whole group indicated that invariance is ensured in scalar level. Hence, item factor loads and factor correlations are similar in both years while item residual variances are different.

3.2. MGCFI Results by Gender

Measurement invariance between genders was checked separately for years 2006 and 2015 in whole group. The resulting goodness of fit indexes are given in Table 5.

Table 5. Goodness of fit indexes by gender in 2006 and 2015

2006	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	435,23	40	0.064	0,031	0,97	0,98	
Metric	457,77	47	0.061	0,035	0,97	0,98	0
Scalar	651,07	63	0,063	0,051	0,97	0,97	-0,01
Strict	662,48	71	0,059	0,051	0,97	0,96	-0,01
2015	X^2	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	1249,37	40	0,11	0,043	0,96	0,97	
Metric	1267,83	47	0,11	0,047	0,97	0,97	0
Scalar	1482,44	63	0,098	0,057	0,97	0,97	0
Strict	1544,86	71	0,093	0,058	0,97	0,97	0

For configural invariance; the goodness of fit indexes in Table 5 demonstrate that all indices fall within the acceptable range in 2006; whereas the RMSEA values are outside such range in 2015. This result suggests that the structure of the model remained unchanged for genders across years. Once configural invariance was ensured as prerequisite of metric invariance, the latter was checked.

In relation with metric invariance, Table 5 indicates acceptable limits for all statistics for year 2006 and 2015; while the statistics except for RMSEA fall within acceptable limits for 2015 (RMSEA<.08; SRMR<.1; TLI>.95; CFI>.95). Examination of ΔCFI refers to positive metric invariance (ΔCFI ≤ 0.01). This finding implies that the relationships between the measured traits and science self-efficacy dimension are similar in both genders. After metric invariance was ensured, the next phase was implemented.

Examination of the values in Table 5 for scalar invariance reveals that the RMSEA, value is outside the acceptable limits for 2015 but the other values are acceptable. When ΔCFI values are examined, it is seen that scalar invariance is provided (ΔCFI ≤ 0.01). The finding reflects invaried item factor loads and item constants across genders. When scalar invariance was deemed acceptable, the last stage was implemented.

With respect to strict factorial invariance, goodness of fit indexes in Table 4 refer to acceptable levels for year 2006. However, in 2015, the statistics except for RMSEA are seen at acceptable limits. The ΔCFI value is found to be within the acceptable range for both 2006 and 2015 (ΔCFI ≤ 0.01). The finding suggests that error variances did not vary between genders in 2006 and 2015.

The analysis of the invariance between gender groups indicated that invariance is ensured in strict invariance level in both 2006 and 2015 implementation with respect to ΔCFI values. Hence, item factor loads, factor correlations and error variances are similar between gender groups in both years. However, the model fit of 2015 seem problematic as RMSEA values are really high in al stages. This may implies that the model in 2015 may be revised.

3.3. MGCFI results in gender subgroups across years

As for the third sub-problem of the research, measurement invariance analyses across years were performed separately in gender subgroups. Indeed model invariance was tested between female students in 2006 and female students in 2015; male students in 20016 and male students in 2015 respectively. The resulting coefficients are given in Table 6.

Table 6. Goodness of fit indexes by gender subgroups between years 2006 and 2015

Female	X²	df	RMSEA	SRMR	TLI	CFI	ΔCFI
Configural	691,09	40	0,084	0,035	0,97	0,98	
Metric	728,01	47	0,079	0,04	0,97	0,98	0
Scalar	1323,66	63	0,092	0,1	0,96	0,95	-0,03
Strict	2333,52	71	0,11	0,084	0,94	0,92	-0,03
Male	**X²**	**df**	**RMSEA**	**SRMR**	**TLI**	**CFI**	**ΔCFI**
Configural	993,51	40	0,1	0,043	0,96	0,97	
Metric	1063,86	47	0,095	0,049	0,96	0,97	0
Scalar	1571,52	63	0,099	0,094	0,96	0,95	-0,02
Strict	2903,2	71	0,12	0,085	0,93	0,91	-0,04

To start with, configural invariance was tested through invariance of the model, factor and items in gender subgroups across years. Examination of the indices in Table 6 show that both subgroups have acceptable values except for RMSEA. This result suggests that the model structure remained unchanged in both male and female subgroups across years.

When the indices in Table 6 are examined in relation with metric invariance, female participants meet the acceptable limits for all statistics, while males are seen to be within the acceptable interval for the values except for RMSEA. Examination ΔCFI reveals that metric invariance is satisfied between 2006 and 2015 both in females and males ($\Delta CFI \leq 0.01$). This finding refers to similar relationship between the measured traits in 2006 and 2015 science self-efficacy dimension in female group as well as the male group. Once metric invariance ensured, the next stage of scalar invariance is checked.

Considering scalar invariance, the fit indices in Table 6 reveal that RMSEA values are outside the acceptable interval both for males and females. ΔCFI values show that scalar invariance is not met between 2006 and 2015 both for females and males ($\Delta CFI > 0.01$). The finding implies that item factor loads were unchaged in gender groups across years, but item constants varied. Due to the lack of scalar invariance, it is questionable to compare gender averages across years. Likewise, strict factorial invariance was not tested.

The analysis of the invariance between implementations within each gender groups indicated that invariance is ensured only in metric level in both male and female group. Hence, item factor loads, factor correlations and error variances are different between 2006 and 20165 implementations in female group as well as male group.

4. DISCUSSION

In this study, it was intended to find out whether science self-efficacy scale of Turkish students taking PISA 2006 and PISA 2015 exams changes across years and gender.

In the model for the self-efficacy scale towards science; it was observed that configural metric and scalar invariance are satisfied for the total group across years; while strict factorial invariance is not satisfied. It was found out that the variables in the model manifest similar factor loads and factor correlations but different error variances across years. On the other hand, separate investigation of invariance by gender subgroups in years 2006 and 2015 revealed ensuring of configural invariance, metric invariance, scalar invariance, and strict factorial invariance. The finding of this study does not seem to be in parallel with findings by Uyar and Doğan (2014) and Gülleroğlu (2017). In neither study above, strict factorial invariance was not provided according to gender group, but it was satisfied in present study. In addition, Uzun and Öğretmen (2010) found out that science self-eficacy variable meets metric invariance according to gender. The finding of Uzun and Öğretmen seems to be at odss with our findings. On the other hand, the results are parallel with findings obtained by Başusta and Gelbal (2015) from a study conducted on PISA 2009 focusing on invariance of the items in science and technology scales across gender. It was found out that the variables in the model revealed similar factor loads, factor correlations and error variances between different gender sub-groups for both years. This suggests that averages of the variables in the science self-efficacy scale can be comparible across gender subgroups. In studies by Saracaloğlu, Yenice and Özden (2013) and Balbağ and Balbağ (2016), it was found out that self-efficacy perception regarding Science and Technology Literacy does not show a significant variance across gender. In this regard, it can be argued that the measurements obtained from the model established with science self-efficacy in the PISA students' survey can be generalized for gender.

In gender subgroups, only configural and metric invariance were satisfied between 2006 and 2015 in both male and female groups, but scalar and strict factorial invariance could not be met. It was found out that the model varibles show similar factor loads between gender groups across years. In other words, the structure was found to be constant across genders. On the other hand, the respondents of 2006 and 2015 applications have divergent interpretations could have an influence on the lack of scalar and strict factorial invariance. During the 9-year period between the two applications, the self-efficacy of different genders towards science could have differed. Still, bearing in mind the probability that the modelled structure cannot remain unchanged, one reason for the lack of invariance could be because the existing structure might have changed. We think that it is needed to write items conforming to current conditions by revising the items in students' questionnaires covered under international tests like PISA guiding national educational policies in the light of application experience. For example, life on Mars did not raise as heated debate as now in 2006. However, today research at NASA is underway for trip to Mars. Considering these developments, we believe that it is no longer possible to interpret item 7 in the questionnaire, which reads as “Discuss how new evidence could change the debate whether life exists on the Mars” in the same way in practice in the course of time. In addition, while the questionnaire was given as a paper-pencil test in 2006, the application became computer-based in 2015. So, computer skills of students may affect their responses, this also could account for the lack of invariance.

Lastly, Bulut, Palma, Rodrigez and Stanke (2015) studied parameter invariance of support and positive identity scales among White and Latin American students across years. They found out that subgroup-year interaction has a significant effect on parameter shift. Hence, the variance of scale parameters between two different groups of students differs across years. Huggings-Manley (2016) stated that item parameters remain unchanged in applications performed at different times; yet, parameter difference could appear between subgroups in the course of time. Thus, further examination of the lower values of goodness of fit indexes in male group across years could shed light onto cases, which are not explained by studies on group invariance and bias. This can also be a sign that self-efficacy pattern varies differently among girls and boys across years. Moreover, recent social responsibility projects promoting sciences for female students may also have had an effect on their self-efficacy. Determining possible causes of these results remain out of the scope of our research. Still, it is advisable to carry out more in-depth research to this end in future studies.

ORCID

Nermin Kıbrıslıoğlu Uysal https://orcid.org/0000-0002-9592-469X
Çiğdem Akın Arıkan https://orcid.org/0000-0001-5255-8792

5. REFERENCES

Akyıldız, M. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6(1)*, 18- 47.

Andrew, S. (1998). Self-efficacy as a predictor of academic performance in science. *Journal of Advanced Nursing, 1998, 27*, 596–603, doi: 10.1046/j.1365-2648.1998.00550.x.

Asil, M. &Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Eğitim ve Bilim, 37(166),* 236-249.

Balbağ, M. Z., & Balbağ, N. L. (2016). Öğretmen adaylarının fen ve teknoloji okuryazarlığına ilişkin özyeterlik algıları ile bilgi okuryazarlıkları arasındaki ilişkinin incelenmesi. *Pegem Atıf İndeksi*, 429-446.

Bandura, A., (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84 (2),* 191-215.

Bandura, A., (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37 (2), 122-147.

Başusta, N. B., & Gelbal, S. (2015). Gruplararası Karşılaştırmalarda Ölçme Değişmezliğinin Test Edilmesi: PISA Öğrenci Anketi Örneği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30(4),* 80-90.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care, 44,* 176-181.

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),* doi: 10.1615/JWomenMinorScienEng.v7.i4.10.

Bulut, O., Palma, J., Rodrigez, M. C. & Stanke,L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *SAGE Open,* 1–18, doi: 10.1177/2158244015586238.

Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi elkitabı*. Ankara: Pegem Yayıncılık.

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, doi: 10.1207/S15328007SEM0902_5

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları.* Pegem Akademi.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5,* 23-35.

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,* 78-94.

Gülleroğlu, H. D. (2017). PISA 2012 Matematik Uygulamasına Katılan Türk Öğrencilerin Duyuşsal Özeliklerinin Cinsiyete Göre Ölçme Değişmezliğinin İncelenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 37(1),* 151-175.

Güzeller, C. (2011). PISA 2009 Öğrenci Anketinde Yer Alan Bilgisayar Tutum Boyutunun Kültürlerarası Eşitliğinin İncelenmesi. *Eğitim ve Bilim*, *36(162),*320-327.

Huggings-Manley, A.C. (2016). Psychometric Consequences of Subpopulation Item Parameter Drift. *Educational and Psychological Measurement, 77(1),* 143–164. doi: 10.1177/0013164416643369

Karakoc Alatli, 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. *Eurasian Journal of Educational Research, 66,* 389-406. http://dx.doi.org/10.14689/ejer.2016.66.22

Kıbrıslıoğlu, N. (2015). *PISA 2012 Matematik Öğrenme Modelinin Kültürlere ve Cinsiyete 476 Göre Ölçme Değişmezliğinin İncelenmesi: Türkiye -Çin (Şangay) -Endonezya Örneği.* Yayınlanmamış yüksek lisans tezi. Hacettepe Üniversitesi

Kline, R.B., (2011). *Principles and Practices of Structural Equation Modelling*. New York, The Guilford Press.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries *Learning and Individual Differences 19,* 355–365.

Lent, R. W., Brown, S. D., & Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology, 33(3),* 265-269.

OECDa (Organisation for Economic Co-operation and Development). http://www.oecd.org/

OECDb (Organisation for Economic Co-operation and Development). http://www.oecd.org/education/

OECD (Organisation for Economic Co-operation and Development) (2015). *PISA 2015 Technical Report.* http://www.oecd.org/pisa/data/2015-technical-report/

Marsh, H. W., Hau, K. T., Artelt, C., Boument, J., & Peschar, J. (2006). OECD's brief selfreport measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6 (4),* 311-360.

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). New York, NY: Guilford Press.

Ölçüoğlu, R., & Çetin, S. (2016). TIMSS 2011 Sekizinci Sınıf Öğrencilerinin Matematik Başarısını Etkileyen Değişkenlerin Bölgelere Göre İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 7(1), 202-220.*

Post, P., Stewart, M. A., & Smith, P. L. (1991). Self-efficacy, interest, and consideration of math/science and non-math/science occupations among Black freshmen. *Journal of Vocational Behavior, 38(2),* 179-186. Doi: https://doi.org/10.1016/0001-8791(91)90025-H

Saracaloğlu, A. S., Yenice, N., & Özden, B. (2013). Fen bilgisi, sosyal bilgiler ve sınıf öğretmeni adaylarının öğretmen öz-yeterlik algılarının ve akademik kontrol odaklarının incelenmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 34(2),* 227-250.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: review of practice and implication. *Human resources management review, 18,* 210-222.

Scott, A. B., & Mallinckrodt, B. (2004). Parental Emotional Support, Science Self-Efficacy, and Choice of Science Major in Undergraduate Women. *The Career Development Quarterly, 53 (3),* 263-273, doi: 10.1002/j.2161-0045.2005.tb00995.x

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003), Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodnessof-Fit Measures, *Methods of Psychological Research Online, 8* (2), pp.23-74.

Tabachnick, B. G., & Fidell, L. S. (2007*). Using Multivariate Statistics* (5. Eds). Boston: Pearson Education.

Uyar. Ş. ve Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2,* 30-43.

Uzun, B., Öğretmen, T. (2010). Fen basarisi ile ilgili bazı değiskenlerin TIMSS-R Türkiye örnekleminde cinsiyete göre ölçme değismezliğinin değerlendirilmesi. *Eğitim ve Bilim, 35(155),* 26-35.

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 (4),* 4-70.

Widaman, K. F., & Reise, S. P., (1997). Exploring the measurement invariance of psychological instruments: Applications in substance use domain. The science of prevention: *Methodological advances from alcohol and substance abuse research,* 281-324.

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.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*, 82-91.