KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: March 17, 2016
Revision received: June 29, 2016
Accepted: July 21, 2016
OnlineFirst: October 20, 2016

www.estp.com.tr
DOI 10.12738/estp.2017.1.0223 ♦ February 2017 ♦ 17(1) ♦ 47–62

Research Article

Testing Measurement Invariance of the Students' Affective Characteristics Model across Gender Sub-Groups

Ergül Demir[1]
Ankara University

Abstract

In this study, the aim was to construct a significant structural measurement model comparing students' affective characteristics with their mathematic achievement. According to this model, the aim was to test the measurement invariances between gender sub-groups hierarchically. This study was conducted as basic and descriptive research. Secondary level analyses were conducted on the Program for International Student Assessment (PISA) 2012 Turkish student questionnaire data. The sample consisted of 4,848 fifteen-year-old students from 170 schools and 12 statistical territories. For analysis techniques, regression analysis, exploratory, and confirmatory factor analyses were executed in order to construct a significant initial measurement model. Multi-group confirmatory factor analysis was executed to analyse the invariance between gender sub-groups. According to the results, taking into consideration the limitations of the model which was constructed in this study, it was observed that strong invariance between gender sub-groups was provided. This finding indicates that there are similarities of affective characteristics for fifteen-year-old Turkish students across gender sub-groups. The results can be evaluated as evidence that the possibility of bias or prejudice in students' affective characteristics toward mathematics is not high.

Keywords

Measurement invariance • Multigroup confirmatory factor analysis • Affective characteristics • PISA 2012 • Gender differences

[1] **Correspondence to:** Ergül Demir (PhD), Department of Measurement and Evaluation, Ankara University, Cebeci Ankara 06590 Turkey. Email: erguldemir@ankara.edu.tr

Citation: Demir, E. (2017). Testing measurement invariance of the students' affective characteristics model across gender sub-groups. *Educational Sciences: Theory & Practice*, *17*, 47–62. http://dx.doi.org/10.12738/estp.2017.1.0223

edam

In the psychological and behavioral sciences, latent variables and traits models are popular subjects of research. The concept of "latent traits" or "latent characteristics" can be observed only by an indirect way, and generally they can be modeled only as very complicated structures. Because these models include many indirectly observed variables, their limitations are high.

It is important that psychological structures meet invariance across sub-groups (like cultural or ethnical groups, gender or territorial sub-groups, etc.). If a structure is invariant, it provides evidence that there is no measurement bias among sub-groups as the systematic error (Little, 1997; Lord, 1980). Also, this is a confirmation in terms of reliability and validity of the results of the measurements (Meredith, 1993; Vandenberg & Lance, 2000). On the other hand, if invariance is not met, and if there is no validity and reliability problem, it indicates that there could be real differences among sub-groups in the limitations of the structure. So, it is obvious that a concept is needed to discuss these psychometric characteristics of the psychometric models. One of these concepts is defined as "measurement invariance" or "measurement equivalence."

Measurement invariance is defined as the equivalency of the latent structure across different groups or sub-groups. When some groups are compared to each other in a latent traits model, the parameters of this model, depending on group memberships, should be the same. Measurement invariance indicates that (a) psychological structures are generalizable to each sub-group; (b) these structures are not affected by subgroup differences, and (c) bias and errors of measurement are minimal. Also, in this context, invariance is defined as the psychometric characteristics of the measurement scale that includes configural invariance, metric invariance, scalar invariance, and residual invariance hierarchically. Configural invariance means that the latent structure model should be the same for each group or sub-group. Metric invariance, also evaluated as "weak invariance," means that item loadings should be the same across groups. This indicates that there is no item bias across groups as a systematic error. Scalar invariance, also evaluated as "strong invariance," means that the vectors of the item intercept should be equivalent across groups. This indicates that there are the same correlations across factors on the model. Residual invariance, also evaluated as "strict invariance," means that items or observed variables of the model should have the same measurement errors. If configural, metric, or scalar invariances are met, this can be evaluated as "partial invariance" (Byrne, 2006; Byrne, Shavelson, & Muthén, 1989; Kline, 2011; Little, 1997; Meredith, 1993).

Confirmatory factor analysis (CFA) and item response theory (IRT) models are reported as two powerful methods for testing the comparability of psychological measurements (Jöreskog & Sörbom, 1996; Kline, 2011; Lord, 1980; Reise, Widaman,

& Pugh, 1993). Between these two techniques, the most common technique to test measurement invariance is "multi-group confirmatory factor analysis (MGCFA)." The baseline of this technique is associated with a covariance structure model developed by Jöreskog (Byrne, 2006; Kaplan, 1995; Little, 1997). Jöreskog (1971) defined the process of testing invariance as a hierarchical set of steps. These steps begin with the construction of a well-fitting multi-group model and this model is commonly known as "factorial or configural model." In this initial model, all parameters remain free across groups and there is no restriction. Then, the parameters of this model (factor loading, error correlations, error variances, and factor variances) are restricted in a logically ordered way. After each restriction, the level of invariance across groups is evaluated. If violations are observed in the first step, that means there is no measurement invariance across groups. Otherwise, it can be stated that "configural invariance" is met. If there is no violation after the restrictions of factor loadings that means "metric invariance" is met, and it can be reported as "weak invariance." If the model remains the same across groups when error correlations of factors on the model are restricted, that means that scalar invariance is met. All these three steps indicate the "partial invariance." If the model remains the same when all parameters are restricted, then "strict invariance" is met across groups. In the MGCFA technique, goodness of fit statistics (like model fit x^2, RMSEA, RMR, CFI, GFI, etc.) and the difference of these statistics for two successive steps are considered for this purpose (Byrne, 2008; Horn & McArdle, 1992; Jöreskog & Sörbom, 1996; Muthén, 1993).

In the related literature, it is seen that most of the studies on measurement invariance, especially initial studies, are very technical and of a theoretic baseline (French & Finch, 2006; Horn & McArdle, 1992; McArdle & Cattle, 1994; Jöreskog, 1971; Koh & Zumbo, 2008; Little, 1997; Meade & Lautenschlager, 2004; Meredith, 1993). In most of these studies, analyses were conducted on simulative data. Also, most of them aimed to improve the technical framework of the concept of invariance and to define the best conditions to implement this technique.

Besides these theoretical studies, the most widely utilized measurement of invariance is for cultural comparisons (Akyıldız, 2009; Berry, Poortinga, Segall, & Dasen, 1992; Wu, Li, & Zumbo, 2007). With the results of these studies, similarities, and differences across cultures are determined and discussed with detail. Also, it is stated that these kinds of results provide us an opportunity to understand the cultural characteristics of groups and countries.

As for other research areas, testing invariance can be used to provide validity evidence for specific measurement tools (Grouzet, Otis, & Pelletier, 2006; Marsh, Hau, Artelt, Baumert, & Peschar, 2006; Zhu, Sun, Chen, & Ennis, 2012). In these studies, it was generally observed that strict invariance was not met in the limitations of the measurement model. Further, it was stated that strict invariance was difficult to provide in psychological and educational structures.

There is some research or parts of research where measurement invariance across gender sub-groups is tested. Hirschfeld and Brown (2009) investigated the structural relationships to achievement in reading comprehension across student sex, year level, and ethnicity in a sample of 3,506 students from New Zealand with MGCFA. According to the results, they observed statistically significant differences for sex, year, and ethnicity. They interpreted the sex differences as the real-world differences in approaches to learning of students. Year level differences were associated with the participation in the New Zealand national qualification assessment system. Ethnic differences were evaluated as bias and prejudice in schooling. Grouzet et al. (2006) examined the measurement invariance of the "Academic Motivation Scale (AMS)" across both gender and time in a longitudinal design. 322 boys and 321 girls in 8th, 9th, and 10th grades completed the French version of the AMS over a 3-year period from 2001 to 2003. According to the results, it was observed that longitudinal cross-gender metric invariance was provided for AMS. This finding was evaluated as weak invariance, and there could be real differences between girls and boys about their academic motivations. Marsh et al. (2006) researched the psychometric characteristics of "Students' Approaches to Learning (SAL)" instruments developed by Organization for Economic Co-operation and Development (OECD) in 4,000 fifteen-year-old students from 25 countries with MGCFA. They found that the factor structures of SAL were well-defined and reasonably invariant across 25 countries. On the other hand, they observed the relations between SAL factors and gender, socioeconomic status, math achievement, and verbal achievement of the fifteen-year-old students. Uzun and Ogretmen (2010) investigated the factors that were related to the fifteen-year-old students' science achievement and assessed the invariance of these factors across gender in a Trend of International Mathematic and Science Study (TIMSS) 2009 sample from Turkey. They found no factors met the strict invariance. It was observed that just partial invariance was met across gender sub-groups. They concluded that comparisons between gender sub-groups were misleading. Uyar and Dogan (2014) examined the invariance of a "learning strategies model" across gender, school type, and territorial subgroups in a PISA 2009 Turkish sample. It was reported that this model provided just weak invariance across gender and school types, and that strong and strict invariance were not met. Further, strict invariance was observed across territorial sub-groups. Başusta and Gelbal (2015) investigated the factors of the PISA student questionnaire items and assesed the measurement invariance of these factors across gender in a PISA 2009 sample from Turkey. As a result, they observed that there were no invariance problems across gender.

All of the research mentioned above indicates that measurement invariance is one of the most important psychometric characteristic of the measurement tools and measurement processes. If we have a tool that provides invariance across groups and this is confirmed, then it is possible that the observed differences indicate real differences. For gender sub-groups, the research shows it is difficult to provide strict invariance. Mostly, just weak or partial invariance levels are provided across gender sub-groups. These findings may not only indicate measurement problems but also gender differences or a gender gap. In education, a gender gap means that students do not have the same opportunities and there could be inequality of opportunity in education. Also, if gender differences are observed on affective characteristics, it is thought that these differences have the potential to explain the differences of cognitive characteristics, like achievement or intelligence. So, affective characteristics would provide the opportunity to understand the nature of cognitive characteristics in indirect way.

In this study, the aim was to construct a significant structural measurement model about students' affective characteristics related with their mathematics achievement by using PISA 2012 student questionnaire data from Turkey. According to this model, the aim was to thoroughly and hierarchically test the measurement invariances across gender sub-groups of the Turkish students. Basically, this main hypothesis was tested: "The students' affective characteristics towards math have the potential to explain the differences between the gender subgroups."

Method

Research Model

This study was conducted as basic and descriptive research. The basic research aimed to further the theorhetical knowledge about the phenomena and variables. In the descriptive research, as one level of the basic research, the aim was to define phenomena and variables as they are (Karasar, 2012; Slavin, 1992). In this study, students' affective characteristics, which are related with their mathematic achievement, were examined according to their gender sub-groups.

Sample

In this study, secondary level analyses were conducted on the PISA 2012 Turkish student questionnaire data. The sample was composed of 4,848 fifteen-year-old students from 170 schools and 12 statistical territories of Turkey. For sampling methods, stratified random sampling was used. The gender of 2,370 of these students (48.9%), is female and the gender of the other 2,478 students (51.1%), is male. There are closed ratios for the distribution of gender.

Data Collection Tool

In this study, secondary level analyses were conducted on the data obtained from the PISA 2012 Turkish student questionnaire. In PISA studies, the student questionnaire is used to gain knowledge about student backgrounds and about their cognitive success or achievement. This tool includes many questions about students' family structure (like parents' educational level, occupational status, wealth, cultural heritage, immigration status, etc.), their educational opportunities (like home possessions, educational sources at home or at school, out of school study time, disciplinary climates, teachers' classroom management, etc.) and their affective characteristics about a domain area (like mathematical anxiety, mathematical behavior, mathematical intentions, mathematical self-concepts, perseverance, etc.) (OECD, 2013, 2014).

Because there are many questions in the PISA student questionnaire, this tool is given to students in three forms (Form A, Form B, and Form C). Each form has around 50 questions. Students are given 30 minutes to answer after beginning the achievement tests. This questionnaire includes common items about family structure and educational opportunities. On the other hand, the affective characteristics of students' and some other related characteristics are questioned with just two forms (OECD, 2013, 2014). So, if affective characteristics are to be studied, there are missing data problems that need to be handled. In order to deal with this problem, it is suggested to use imputation methods like EM algorithms or multiple imputations, because each form is given to students randomly and there is no bias for the missing data mechanism (Allison, 2002; Enders, 2010; Little & Rubin, 1987).

Data Analyses

To begin this study, a structural model was constructed that represented the students' affective characteristics related to their mathematic achievement. For this process, respectively, a stepwise regression analysis, a principal component analysis with oblique rotation as the exploratory factor analysis, and a confirmatory factor analysis were conducted on the PISA 2012 Turkish student questionnaire data.

After defining the model, the measurement invariance of this model across the gender sub-groups of the students was analyzed. For this process, multi-group confirmatory factor analysis (MGCFA) was used. In this technique, measurement invariance is defined with four or five hierarchical steps. When the structural model includes just the first level latent variables, measurement invariance can be analyzed with a four-step process (Jöreskog & Sörbom, 1996). So, in this study, measurement invariance was analyzed by considering the four-step process. These steps are defined as (1) configural invariance, (2) metric invariance, (3) scalar invariance, and (4) strict invariance. For criteria to provide invariance, the hierarchic differences of model-data fit indices (RMSEA, RMR, CFI, GFI, NFI, and NNFI) and the differences of model-data fit x^2 statistics between the steps were considered. When the differences of the model-data fit indices were more than 0.01 and/or x^2 statistics were statistically significant ($p < .05$), these findings were interpreted as a violation of invariance. Otherwise, it was decided that measurement invariance was provided across sub-groups. If there were some violations, the causes and resources of this violation were explained and discussed with a deep analysis based on the differences of the sub-groups' model coefficients.

Before all analyses, the principal assumptions of the analyses were tested carefully. In this context, missing data, extremes, univariate normality, multivariate normality, linearity, multicollinearity, and autocorrelations were tested. Because the sample was very large, normality assumptions were checked by graphical methods instead of hypothesis tests or descriptive statistical calculations. All these results are explained before the findings of each research question in the following "Results" sub-sections.

Findings

In this section, in accordance with the aims, first the construction process of students' affective characteristics model is explained and findings obtained from this model are interpreted. Then findings about invariances across gender sub-groups are reported and criticized separately.

Students' Affective Characteristics toward Mathematics Model

In this study, first a significant measurement model was needed to complete further analyses. In order to construct this model, the variables in the PISA 2012 Turkish student questionnaire data set were examined in detail. It was found that there were more than 80 index variables in this data set. These index variables represent students' characteristics as total standard scores and most of them are continuous. Among these indexes, some of them were not available for the Turkish sample (like cultural heritage, language at home, immigration status, information for the Labour Market, etc.) and there were no data for these variables. So, just 48 of them could be defined for the Turkish sample. Among these 48 index variables, just 12 of them were about students' affective characteristics so could be related with their mathematical achievement.

After that, in order to provide statistical evidence for the relationship between affective characteristics and mathematical achievement, a regression analysis was conducted with these 12 continuous index variables and these variables were defined as predictors. A stepwise model was used for this purpose. After eight steps, a significant regression model was constructed with eight index variables. To test the univariate and multivariate normality with graphical methods, scatter-dot plots were prepared and examined for each variable and each combination of these variables. It was seen that the distributions of the variables were very close to the normal distributions. So, it was decided that there was no violations in normality. According to collinearity diagnostics, there was no multicollinearity problem. Correlations between variables were low but significant at a significance level of 0.01. Values of tolerance were between 0.261 and 0.916. Values of variance inflation factor were between 1.092 and 3.834. Values of condition index (CI) were under 30 and between 1.000 and 4.433. According to the Durbin-Watson statistic (1.068), there was no autocorrelation problem. This model was statistically significant (F = 208.189, df(regression) = 9, df(residual) = 4839, df(total) = 4847, and $p < .001$) and it had 27.8% as the determination level (R = 0.528, R^2 = 0.278, and Adj.R^2 = 0.277). Regression coefficients for each predictor for last step model is shown at the following Table 1.

Table 1

Regression Coefficients for Affective Predictors of Students' Mathematical Achievement

Predictors	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	B		
(Constant)	465.476	1.698		274.105	0.000
Mathematical Self-Efficacy	55.301	1.763	0.489	31.369	0.000
Mathematical Anxiety	-17.371	1.680	-0.159	-10.338	0.000
Mathematical Behaviour	-15.862	1.720	-0.150	-9.224	0.000
Mathematical Work Ethic	-16.875	1.692	-0.184	-9.975	0.000
Mathematical Intentions	5.807	1.513	0.052	3.839	0.000
Perseverance	4.59	1.413	0.045	3.249	0.001
Mathematical Self-Concept	7.713	2.541	0.073	3.036	0.002
Attributions to Failure in Mathematics	-2.799	1.336	-0.027	-2.095	0.036

*Dependent variable: Plausible ,value of mathematics achievement (PV_Math)

As seen in Table 1, the most predictive variable for mathematic achievement is mathematic self-efficacy (β = 0.489). This characteristic is positively correlated with mathematic achievement. It is followed by mathematic work ethic (β = −0.184,) mathematic anxiety (β = −0,159) and mathematic behavior (β = −0.150) respectively. These 4 characteristics are negatively correlated with mathematic achievement.

After defining the predictors of mathematic achievement, 8 continuous index variables were considered to construct a measurement model. Exploratory and confirmatory factor analyses were conducted with these index variables respectively. As the exploratory factor analysis, a principal component analysis with oblique rotation was used as the factoring model. Tabachnick and Fidel (2007) stated that the oblique rotation was more preferable than the other rotation methods if the variables were desired to be kept.

According to the results of exploratory factor analysis, first, it was seen that the data was appropriate for factor analysis (KMO = 0.806, Bartlett's test approximate-x^2 = 12854.069, df = 28, and $p < .001$). Communalities of the index variables were between 0.272 and 0.827. There was a significant structure with two factors. The first factor had 43.298% variance explained. The second factor had 13.839% variance explained. The total variance explained was 57.137%. There was negative correlation between these two factors (r = −0.268 and $p < .05$). The structure matrix after oblique rotation is shown at the following Table 2.

Table 2

Structure Matrix for Eight Index Variables

Variables	Component	
	1	2
Mathematical Self-Concept	.896	
Mathematical Work Ethic	.837	
Mathematical Behaviour	.752	
Mathematical Self-Efficacy	.733	
Perseverance	.543	
Mathematical Intentions	.515	
Attributions to Failure in Mathematics		.835
Mathematical Anxiety		.698

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

As seen in Table 2, the first factor includes six index variables. These variables are expected to be positively correlated with achievement according to the general literature. So, this factor was named "positive characteristics." The second factor includes just two index variables. These variables are expected to be negatively correlated with achievement. So, this factor was named "negative characteristics."

After the exploratory analysis, a confirmatory factor analysis was conducted with these two factors and eight index variables model. Structural equation modeling was used for this purpose. After this, as in exploratory studies, univariate and multivariate normality were tested with graphical methods with these eight variables. For each variable and each combination of these variables, scatter-dot plots were prepared and examined. It was seen that the distributions of the plots were very close to the normal distributions. Also, to make the model-data fit better, some modifications among error covariance were applied to the model. Model-data fit indexes and model-data fit x^2 statistics obtained from the analysis are shown at Table 3.

Table 3
Goodness of Fit Statistics for Students' Affective Characteristics Model

Goodness of Fit Statistics	
Weighted least squares x^2	251.08 (df = 15 and p = .000)
Normed Fit Index (NFI)	0.99
Non-Normed Fit Index (NNFI)	0.97
Comparative Fit Index (CFI)	0.99
Goodness of Fit Index (GFI)	0.99
Adjusted Goodness of Fit Index (AGFI)	0.97
Root Mean Square Error of Approximation (RMSEA)	0.057
Root Mean Square Residual (RMR)	0.021
Standardized RMR	0.027

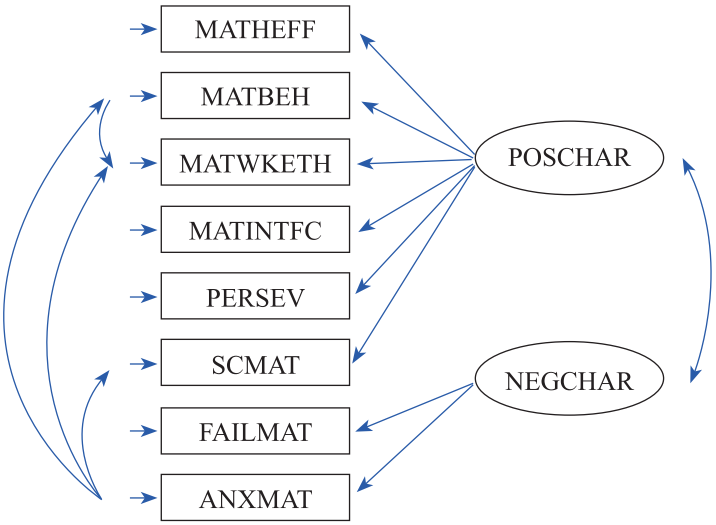
As seen at Table 3, x^2 is statistically significant. Values of NFI, NNFI, CFI, GFI and AGFI are above 0.95. Values of RMR, and SRMR under 0.05 and RMSEA are around 0.05. This goodness of fit statistics point out a perfect model-data fit. Also, raw and standardized coefficients of paths on the model and their t-values are shown in Table 4.

Table 4
Path Coefficients for Students' Affective Characteristics Model

Latent Variables	Observed Variables	Raw Coefficients		Standardized Coefficients		t	*p*
		Error	Path	Error	Path		
Positive Characteristics	Mathematical Self-Concept	0.38	0.53	0.58	0.65	48.44	.011
	Mathematical Work Ethic	0.44	0.56	0.59	0.64	47.00	.012
	Mathematical Behaviour	0.41	0.77	0.41	0.77	59.50	.013
	Mathematical Self-Efficacy	0.55	0.36	0.81	0.43	30.29	.012
	Perseverance	0.64	0.43	0.77	0.48	33.52	.013
	Mathematical Intentions	0.10	0.80	0.14	0.93	77.56	.010
Negative Characteristics	Attributions to Failure in Mathematics	0.69	0.29	0.89	0.32	16.74	.017
	Mathematics Anxiety	0.32	0.62	0.46	0.74	24.67	.025

As seen in Table 4, all path coefficients are statistically significant at a significance level of .05. All error terms are under .90. According to the standardized path coefficients, the best predictor of students' positive affective characteristics is "mathematic intentions." It is followed by "mathematic behavior." For negative characteristics, the best predictor is "mathematic anxiety." The positive and negative characteristics are related each other with a strong negative correlation (r = −0.58 and p = .02).

After all these studies, a statistically significant first level structural model that represented the students' affective characteristics related with their mathematic achievement could be structured and determined. This model includes 8 index variables with two factors and shows a perfect model-data fit. This model was named "students' affective characteristics model" and is shown in Graphic 1. For further measurement, invariances across gender sub-groups were conducted on this model.



Graphic 1. Students' affective characteristics model for PISA 2012 Turkish sample.

As seen in Graphic 1, one of these two factors includes just two variables. As a general guide, it is recommended that a factor should have at least three variables in order to define a factor. But this depends on the design of the study and characteristics of the variables. If some factors have two variables, these factors should be interpreted with caution. This is also possible when the variables are highly correlated with each other and almost uncorrelated with other variables (Kline, 2011; Tabachnick & Fidell, 2007). Indeed, in this study, correlations between variables can be interpreted as high by considering the sample size and significance level ($r > 0.70$ and $p < .001$). Also, the factors are almost uncorrelated with each other. Therefore, although the data included two variables, it was preferred to define them as separate factors and to include them into the model.

Measurement Invariance for Gender Sub-Groups

Measurement invariance across students' gender sub-groups was analyzed with a four-step multi-group confirmatory factor analysis. In each step, goodness of fit statistics obtained from analyses is shown in Table 5.

Table 5
Goodness of Fit Statistics Obtained From Multi-Group Structural Equation Model Analysis for Gender Sub-Groups

Invariance Steps	Gender	Group Goodness of Fit Statistics			Model Goodness of Fit Statistics				x^2	df	*p*
		GFI	RMR	x^2 Contribution (%)	NFI	NNFI	CFI	RMSEA			
Configural Invariance	Female	0.99	0.022	38.88	0.99	0.98	0.99	0.055	262.09	30	.00
	Male	0.98	0.025	61.12							
Metric Invariance	Female	0.99	0.022	39.23	0.99	0.98	0.99	0.049	271.15	38	.00
	Male	0.98	0.028	60.77							
Scalar Invariance	Female	0.99	0.021	39.41	0.98	0.98	0.99	0.050	278.46	39	.00
	Male	0.98	0.028	60.59							
Strict Invariance	Female	0.98	0.030	43.57	0.98	0.98	0.98	0.053	371.73	47	.00
	Male	0.98	0.035	56.43							

As seen in Table 5, Students' Affective Characteristics Model provides structural invariance at the first step, because all goodness of fit statistics are between the acceptable score range for perfect model-data fit. GFI, NFI, NNFI, and CFI are above 0.95. RMR is under 0.05 and RMSEA are around 0.05. Also, model-data fit x^2 value is statistically significant at the significance level of 0.01. These findings point out that the Students' Affective Characteristics Model is significant and available in each gender sub-groups.

As for the second step, it is observed that the model provides metric invariance across gender sub-groups. The differences of the goodness of fit statistics (ΔGFI, ΔRMR, ΔNFI, ΔNNFI, ΔCFI, and ΔRMSEA) are under 0.01. Also the difference of x^2 statistics is not significant ($\Delta x^2 = 9.06$, $\Delta df = 8$, $p > .10$). According to this result, each index variable has the same predictive level and the same order for each gender sub-group. So it could not be observed that there was item bias among gender sub-groups.

Similarly, in the third step, it is observed that the model provides scalar invariance across gender sub-groups. The differences of the goodness of fit statistics (ΔGFI, ΔRMR, ΔNFI, ΔNNFI, ΔCFI, and ΔRMSEA) are under 0.01. Also the difference of the x^2 statistics is not significant ($\Delta x^2 = 7.31$, $\Delta df = 1$, $p > .005$). According to this result, correlations between factors in each sub-group are the same. For both female and male sub-groups, there is significant and strong correlations between the two factors ($r_{female} = -0.67$ and $r_{male} = -0.51$ and $p < .001$).

On the other hand, in the fourth step, even if the differences of the goodness of fit statistics (ΔGFI, ΔRMR, ΔNFI, ΔNNFI, ΔCFI, and ΔRMSEA) are under 0.01, the difference of the x^2 statistics is significant ($\Delta x^2 = 93,27$, $\Delta df = 8$, $p < .005$). It is observed that there are some differences with the rank of index variables according to their error terms. For example, while FAILMAT has the highest error terms for boys, PERSEV has the highest error terms for girls. So it can be stated that strict invariance cannot be provided.

Discussion and Conclusion

One obvious result of this study is that it is very difficult to construct a significant multilevel model which includes relations between affective and cognitive characteristics in education. In this study, such kind of model could be constructed under two factors: with just eight index variables among more than 80 variables. This model was named "students' affective characteristics model." This model provided moderate model-data fit at first. It needed some modifications between the error terms of the observed variables in order to improve the model.

Meredith (1993), Horn and McArdle (1992), and also Byrne (2008) emphasized the importance of the initial model for structural analysis. Constructing a significant initial model was evaluated as the most important and difficult part of the process. In the concept of measurement invariance, there is little difference between the concept of "factorial invariance" and "structural invariance." Factorial invariance indicates that an initial model has statistically significant model-data fit in each group. Further, structural invariance includes other steps; metric invariance, scalar invariance, and strict invariance. In some research, it was reported that one of the possible reasons of providing just partial invariance was the weakness of the initial model or some related measurement limitations (Grouzet et al., 2006; Marsh et al., 2006; Uzun & Ogretmen, 2010; Zhu et al., 2012).

As another result of this study, it was observed that the students' affective characteristics model provided not strict invariance, but strong invariance across gender sub-groups in the PISA 2012 Turkish sample. This model was partially invariant across sub-groups. This result is supported by some other research (Grouzet et al., 2006; Hirschfeld & Brown, 2009; Uyar & Dogan 2014; Uzun & Ogretmen, 2010). In these studies, affective and cognitive models were tested and it was observed that just weak or partial invariance would be provided across gender groups. Mostly, these findings were evaluated by associating them with real-gender differences in real life. On the other hand, Başusta and Gelbal (2010) observed that there was no invariance problem in the PISA 2009 student questionnaire items across gender groups. It is thought that this different finding arises from the evaluation criteria used for the significance of invariance. The researchers considered the changes of CFI indexes in two successive stages instead of the changes of x^2. In this study, the chances of x^2 were also considered to evaluate the invariance.

In this study, the partial invariance of the model across groups indicates that there are similarities of affective characteristics for fifteen-year-old Turkish students across gender sub-groups. So this can be evaluated as evidence that the possibility of bias or prejudice in students' affective characteristics toward mathematics is not high. It is known that there could be significant differences in students' cognitive characteristics, like academic achievement or intelligence, across gender sub-groups. Also, there are some differences in the affective characteristics (MEB, 2015; OECD, 2015a, 2015b). However, it is understand that the possibility of observing gender differences or a gender gap in affective characteristics is lower than for cognitive characteristics. On the other hand, if the affective characteristics model did not provide invariance across gender sub-groups, it would indicate that the differences of students' achievements between gender sub-groups could be explained by the differences in cognitive characteristics. Finally, according to the results of this study, it was seen that the potential of the affective characteristics in order to explain the students' achievement is low in the PISA 2012 Turkish sample.

References

Akyıldız, M. (2009). The comparison of construct validities of the PIRLS 2001 test between countries. *Yüzüncü Yıl University Journal of Education, 6*(1), 18–47.

Allison, P. D. (2002). *Missing data.* California, CA: Sage.

Başusta, N. B., & Gelbal, S. (2015). Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire. *Hacettepe University Journal of Education, 30*(4), 80–90.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology: Research and applications.* Cambridge, MA: Cambridge University Press.

Byrne, B. M. (2006). *Structural equation with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associations.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872–882.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: The Guilford Press.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling-A Multidisciplinary Journal*, *13*(3), 378–402. http://dx.doi.org/10.1207/s15328007sem1303_3

Grouzet, F. M. E., Otis, N., & Pelletier, L. G. (2006). Longitudinal cross-gender factorial invariance of the Academic Motivation Scale. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 73–98. http://dx.doi.org/10.1207/s15328007sem1301_4

Hirschfeld, G. H. F., & Brown, G. T. L. (2009). Students' conceptions of assessment factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment*, *25*(1), 30–38. http://dx.doi.org/10.1027/1015-5759.25.1.30

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement equivalence in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modelling: Concepts, issues and applications* (pp. 100–117). Thousand Oaks, CA: Sage.

Karasar, N. (2012). *Bilimsel araştırma yöntemi* (24th ed.) [Scientific research methods]. Ankara, Turkey: Nobel Akademik Yayıncılık.

Kline, R. B. (2011). *Principles and practices of structural equation modelling*. New York, NY: The Guilford Press.

Koh, K. H., & Zumbo, B. D. (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.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53–76.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Marsh, H., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*(4), 311–360. http://dx.doi.org/10.1207/s15327574ijt0604_1

McArdle, J. J., & Cattel, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research, 29*(1), 63–113.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*(1), 60–72.

Meredith, W. (1993). Measurement invariance, factor analysis and factoral invariance. *Pcychometrica, 58*(4), 525–543.

Milli Eğitim Bakanlığı. (2015). *PISA 2012 araştırması ulusal nihai raporu* [PISA 2012 national final report]. Ankara, Turkey: Author.

Organisation for Economic Cooperation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en_

Organisation for Economic Cooperation and Development. (2014). *PISA 2012 technical report.* PISA, OECD Publishing. Retrieved March 15, 2016 from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Cooperation and Development. (2015a). *Education at a glance 2015: OECD indicators*. OECD Publishing. http://dx.doi.org/10.1787/eag-2015-en

Organisation for Economic Cooperation and Development. (2015b). *The ABC of gender equality in education: Aptitude, behaviour, confidence.* PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264229945-en

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor-analysis and item response theory - 2 approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. http://dx.doi.org/10.1037//0033-2909.114.3.552

Slavin, R. E. (1992). *Research methods in education* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Uyar, Ş., & Dogan, N. (2014). An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample. *International Journal of Turkish Education Sciences, 2*(3), 30–43.

Uzun, B., & Ogretmen, T. (2010). Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey. *Education and Science, 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*(1), 4–69.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research & Evaluation, 12*(3), 1–26.

Zhu, X., Sun, H., Chen, A., & Ennis, C. (2012). Measurement invariance of expectancy-value questionnaire in physical education. *Measurement in Physical Education and Exercise Science, 16*(1), 41–54. http://dx.doi.org/10.1080/1091367X.2012.639629