ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(2); 154-170

Investigation of Measurement Invariance of Turkish Subtest on ABIDE 2016 in Relation to Characteristics of Teachers: Sub-sampling Method

Süleyman ÜLKÜ*   Burcu ATAR **

Abstract

The Ministry of National Education carried out the ABIDE (Monitoring and Evaluation of Academic Skills) in 2016 in order to test the knowledge and skills of 8th-grade students. Since the ABIDE 2016 study was implemented for the first time in our country, it is very important to prove measurement invariance for the validity of the results. Within the scope of this research, the measurement invariance of the success of the students in the Turkish test according to the education level and professional experience of the teachers was examined. In the research, data were obtained from the Ministry of National Education, Directorate-General of Measurement, Evaluation, and Examination Services. Responses of students to the multiple-choice items in the ABIDE 2016 Turkish test and teacher questionnaire data were used in the study. All the data were used in the investigation of measurement invariance according to professional experience. Investigation of measurement invariance according to education level was carried out both by using and not using the method of sub-sampling. Factor 10 and Mplus 7 programs were used in the analysis of the data. At the end of the study, the Turkish achievement model provided all levels of measurement invariance among the student groups formed according to the professional experience and education level of the teachers.

Keywords: Measurement invariance, ABIDE 2016, sub-sampling method

Introduction

Education has become one of the globally significant indicators for the attainment of development-focused strategic objectives of countries in recent years. It is possible to forecast the future of a given country based on the effectiveness of educational reforms and the actual student achievement rates. Therefore, standard measurement and evaluation systems are required to evaluate the quality of learning experiences and to provide stakeholders with feedback according to these evaluations.

Exams on an international scale such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are carried out by international organizations in order to provide feedback for the development and improvement of countries' education systems. Similarly, in Türkiye, it was aimed to develop ABIDE (Monitoring and Evaluation of Academic Skills) which is a standard measurement and evaluation tool by the Ministry of National Education (MEB in Turkish). The overall aim of ABIDE is to determine to what extent 8th-grade students have high-level mental skills in Turkish, mathematics, science, and social studies and to reveal the student, family, teacher, and school characteristics that affect the success of the students (MEB, 2017). ABIDE implemented in two-year periods was first implemented at the 8th grade level in 2016, and the second application was made at the 4th and 8th grade levels in 2018. ABIDE, which was planned to be held in 2020, was carried out in 2021 due to the covid-19 pandemic. However, final reports haven't been published by the MEB except for 2016.

The present study is a part of Master's Thesis conducted under the supervision of Prof. Dr. Burcu ATAR and prepared by Süleyman ÜLKÜ

* PhD. Student, Hacettepe University, Faculty of Education, Ankara-Türkiye, suleymanulku@ hacettepe.edu.tr, ORCID ID: 0000-0003-1965-0671
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Ülkü, S., & Atar, B. (2023). Investigation of measurement invariance of Turkish subtest on ABIDE 2016 in relation to characteristics of teachers: Sub-sampling method. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 154-170. https://doi.org/10.21031/epod.1084985

Received: 9.03.2022
Accepted: 5.06.2023

In the ABIDE 2017 report, comparisons of students' achievements are given according to teachers' educational backgrounds and professional experience. Of course, there are many factors such as school, family, teacher, and environment that underlie the success differences of students. The characteristics one must possess have started to differ gradually in recent years; education plays the most vital role in the indoctrination of such characteristics. Thus, teachers assume considerable responsibility in this process. ABIDE results allow for the interpretation of student achievement rates in terms of the characteristics of teachers enabling the institutions concerned and stakeholders in education to adopt measures and take decisions regarding the improvement of the education system (MEB, 2017). In this case, the report outcomes are expected to be valid and reliable among different groups.

In educational studies, comparisons between groups are frequently made to identify the qualities stemming from the individual, school, teacher, etc. affecting student achievement. However, before commenting that "differences between groups stem from variables originating from students, teachers or schools" based on such comparisons, it is necessary to examine whether these differences are caused by the measurement tool or not. In order to make comparisons according to groups using a measurement tool, measurement invariance must first be ensured in those groups, if measurement invariance is not ensured, the results of the comparison will lose their significance (Byrne Barbara, 2004). Measurement invariance denotes testing whether the measurement tool shows a similar structure among different groups to provide evidence for the validity of measurement tools (Van de Schoot, Lugtig & Hox, 2012). In other words, measurement invariance is to obtain similar results by applying the same scale to different groups that are similar in terms of measured characteristics (Cheug and Rensvold, 1998).

The ABIDE 2017 report frequently compares the achievement levels of students from different groups in terms of teacher qualities.

Individuals in different groups yet equivalent in terms of the attributes assessed must obtain the same score for the accuracy of the comparisons (Schmith & Kuljanin, 2008). This means that some evidence must be provided regarding the measurement of similar structures among groups assessed in terms of the attribute assessed. In other words, the measurement invariance of the tests in the groups determined must be established in order to make comparisons among varying groups using the observed variable scores (Vandenberg & Lance, 2000). Explaining the differences in the results obtained from a measurement tool solely based on individual properties in research studies, making comparisons among groups in terms of the variables to be assessed might not always be accurate. This is because the difference among individuals may also result from the measurement tool (Cheug ve Rensvold, 1998).

Obtaining information about the equivalence of the construct validity of the tests given within the scope of the ABIDE evaluation in 2016 among the student groups formed in terms of the education level and professional experience of teachers would contribute to provide the validity of the measurement results.

Measurement Invariance

The concept of validity is defined as supporting the outcomes based on the scores obtained from the measurement tool and the interpretations made with reference to these outcomes with experimental and theoretical evidence (AERA, APA & NCME, 2014). If one is to compare certain structures among various groups with a measurement tool, the theoretical structure must be the same and be interpreted in the same way by the sub-groups. Otherwise, test bias occurs (Kline, 2011). Based on this point, measurement invariance studies are conducted to identify whether a sub-group has an advantage over others or whether the measurement tools show the same structure as the sub-groups. Making comparisons among groups not displaying the same structure causes the measurement tool not to function. This leads to misinterpretations, which gives rise to misjudgments.

There are various measurement invariance analysis methods in the existing body of literature. The first group of methods examines differences in item and test functions based on the Item Response Theory, the second group consists of methods based on Latent Class Analysis and the final group includes the methods of multi-group confirmatory factor analysis (MGCFA) based on structural equation modeling and the invariance of mean and covariance structures (Kankaras et. al, 2011). MGCFA testing the equivalence of covariance structures is frequently used in measurement invariance studies (Meredith, 1993).

Structural Equation Modeling (SEM) is a powerful and advanced statistical tool providing the researcher with a comprehensive method to assess and modify the model created through theoretical inferences (Dragan & Topolsek, 2014). According to Tabachnick & Fidel (2013), structural equation modeling (SEM) denotes a collection of statistical techniques allowing for the examination of the relationships between one or more independent variables, either continuous or discrete, and one or more dependent variables, either continuous or discrete. SEM analyses signify an expanded combination of factor analysis, multiple regression, and covariance analysis (Hoyle, 2012).

According to Kline (2011), SEM involves six steps: model specification, model identification, evaluation of model fit, measurement of fitness statistics, re-specification of the model where necessary, and reporting of the results. A frequently employed method in SEM analyses, MGCFA is a technique requiring the simultaneous application of CFA on two or more groups. This analysis tests whether the model created by the researcher for the measurement tool is the same for the sub-groups of the sample (Tabachnick & Fidell, 2013).

According to Vanderberg & Lance (2000), measurement invariance is handled by multi-group confirmatory factor analysis as follows: Let us assume the score obtained by the individual i within the group k for the assessed variable of j is X_{ijk}. In this case, the factor model for X_{ijk} is as follows.

$$X_{ijk} = \tau_{jk} + \Upsilon_{jk} W_{jk} + u_{jk} \tag{1}$$

τ_{jk} represents the coefficient factor between the observed and latent structure, Υ_{jk} signifies the factor loadings matrix of rx1 considering that r represents the number of items, W_{jk} shows the common factor loadings vector matrix for i individuals in the rx1 pattern, and u_{jk} shows the error vector of independently observed variables. Furthermore, j signifies the assessed variable, k the group, and i the individual. In this case, X_{ijk} is signified as the score of the individual i within the group k for the variable j. Additionally, it is assumed that the measurement errors are within themselves and the correlation with the common factor loadings is "0". Based on the assumption E(W_{jk}, u_{jk})=0, the covariance equation is as follows:

$$cov(X_{ijk}) = \Sigma_k = \Lambda_k \Phi_k \Lambda'_k + \theta_k \tag{2}$$

The expression Λ_k signifies the matrix of the pxr pattern consisting of Υ_{jk} while Φ_k indicates the variances and covariances in Υ_{jk}. θ_k signifies the diagonal matrix of measurement errors. Similarly, the average vector of X_{ik} can be expressed as follows:

$$E(X_{ik}) = \mu_k = \tau_k + \Lambda_k K_k \tag{3}$$

Based on the equations given above, the question of whether the parameters of $[\tau_k, \Lambda_k,\ \theta_k]$ are equal in k groups (Vandenberg & Lance, 2000, p. 10; Jöreskog & Sörborm, 1993; Millsap & Olivera-Aguilar, 2012, p. 381).

Measurement invariance is exhibited with multi-group confirmatory factor analysis through the testing of the four nested hierarchical levels or the hypothesis. These four levels are called configural, metric, scalar, and strict invariance, respectively (Meredith, 1993).

Configural Invariance

According to Wu, Li & Zumbo (2007), it denotes the initial level of measurement invariance analysis and constitutes a prerequisite for continuing with other levels. This level involves testing whether the model (factor structure) established based on the research hypothesis is the same among the groups. In other words, it means that the Λ_k the matrix in equation 3 has the same fixed and free factor loads for all groups (Widamann ve Reise, 1997).

$$\Lambda_k{}^{(1)} = \Lambda_k{}^{(2)} \tag{4}$$

If configural invariance is not ensured, in other words, if factor structure is the same among groups, the factor configuration among the groups does not differ and the items measure the same structure among different groups. If configural invariance is not ensured, there is no need to conduct the analyses to identify the differences among groups or test the remaining levels of measurement invariance as the measured configurations differ from one group to another (Vandenberg & Lance, 2000).

Metric Invariance

This invariance level is also called weak invariance (Meredith, 1993). In addition to configural invariance, metric invariance is based on the condition whereby the factor loadings of the items concerned must be equal among the groups.

$$\Upsilon_{jk}{}^{(1)} = \Upsilon_{jk}{}^{(2)} \tag{5}$$

Observed variables are connected to latent variables through factor loadings. Therefore, even a minute change in the latent variable affects the observed variable (Bollen, 1989). For this reason, factor loadings must be equivalent if one wants to measure the same configuration among different groups.

The fitness of the metric model is compared to that of the structural model using the difference between the chi-square tests or fit indices to identify whether the condition of metric invariance is fulfilled. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the factor loadings in the sub-groups subject to the comparison do not change. In this respect, this means that all individuals in the sub-groups interpreted the items similarly. The factor variances and covariances may be compared among the groups with the fulfillment of the condition of metric invariance. However, it is not possible to indicate exactly the source of the average difference among the groups.

If the condition of metric invariance is not fulfilled, one might indicate that factor loadings vary among the groups and people made different interpretations of the items concerned (Bialosiewicz, Murphy & Berry, 2013). The lack of metric invariance may signify that the meanings of the items are not the same for all groups, leading to item bias. Partial measurement invariance studies may be conducted if this is the case. If the condition of metric invariance is satisfied, one might move to the next level.

Scalar Invariance

In addition to the conditions required by metric invariance, scalar invariance is based on the equivalence of item threshold values for the sub-groups.

$$\tau_{jk}{}^{(1)} = \tau_{jk}{}^{(2)} \tag{6}$$

To assess scalar invariance, the fitness of the model established is compared with that of the metric model by using the difference between the chi-square difference tests or fit indices. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the factor threshold values do not vary among the sub-groups (Vandenberg & Lance, 2000).

The fulfillment of the condition of scalar invariance means that the averages of factors and observed variables may be compared. In other words, one might argue that there is no bias favoring any sub-group(s) and that the average differences in observed variables source from those in the latent variable (Başusta & Gelbal, 2015). Strict invariance is the next level following the fulfillment of the condition of scalar invariance.

Strict Invariance

At this level, the condition taken into consideration in addition to the conditions of scalar invariance is the equivalence of item error variances among the sub-groups.

$$\theta_k{}^{(1)} = \theta_k{}^{(2)} \tag{7}$$

To assess strict invariance, the fitness of the model established is compared with that of the model established at the level of scalar invariance by using the difference between the chi-square difference tests or fit indices. If there are no significant differences in model fit or if the fit indices are within the desired range, one might argue that the item error variances do not vary among the sub-groups (Bollen, 1989).

If the condition of strict invariance is fulfilled, one can compare observed variances and covariances in addition to the averages of latent and observed variables. However, one must also keep in mind that strict invariance is a quite limited model, and its conditions are rarely fulfilled in practice. This is because as the variance resulting from the latent variable increases, so do the residual variances of the items (Bialosiewicz, Murphy & Berry, 2013).

Purpose of Study

The present study examines whether measurement invariance is established among student groups created based on the education level and professional experience of teachers for the ABIDE 2016 Turkish test. In this respect, the following research questions were identified: (a) "Is measurement invariance established among student groups formed on the basis of the professional experience of teachers in the ABIDE 2016 Turkish test?", (b) "Is measurement invariance established among student groups formed on the basis of the education levels of teachers using the sub-sampling method in the ABIDE 2016 Turkish test?", and (c) "Is measurement invariance established among student groups formed on the basis of the education level of teachers without using the sub-sampling method in the ABIDE 2016 Turkish test?"

Method

The study is a descriptive research in order to illuminate a given situation and to determine the level of validity of the study, which aims to examine the measurement invariance of students' success in Turkish tests according to teachers' education level and professional experience. Studies that aim to reveal a situation without intervening are in the type of descriptive research (Fraenkel & Wallen, 2006; Karasar, 2011). Descriptive models are research models that aim to reveal the states of variables and to reveal the change between variables (Gall et. al, 1999).

Research Population and Sample

The population of the ABIDE assessment consists of 8th-grade students from Türkiye. The ABIDE 2016 assessment was applied in 16,118 schools and 48,091 classes. Conducted in all 81 provinces, the study took into consideration around 400 students from each province. The number of students to be included in the samples in metropolises was increased proportionately to the population to better reflect the overall population. Therefore, the assessment was given to a total of around 38,000 students. Furthermore, students were also classified into stratas through stratified sampling in order for the samples selected to better represent the province concerned (MEB, 2017).

For research purposes, the data on 7952 students using Form A of the Turkish test were obtained from the Directorate-General of Measurement, Evaluation, and Examination Services. 86 students not providing answers to any questions in Form A of the Turkish test were excluded from the study. As a result of the examination of missing values, data concerning a total of 365 students were excluded from the study. As a result, data from 7501 students were used in the analyses. Table 1 shows the information about the student groups created based on the professional experience and education levels of teachers included in the research sample.

Table 1

Student Frequencies in Terms of the Professional Experience and Education Status of Teachers

Professional Experience of Teachers	N of Students	Education Status of Teachers	N of Students
0-5 (short)	2.183	Associate Degree	240
6-15 (medium)	3.790	Bachelor's Degree	6.997
16+ (extensive)	1.528	Master's Degree	264
Total	7.501	Total	7.501

Upon forming groups based on the educational background of teachers within the scope of the study, great differences seemed to emerge among the student groups. Within this scope, the study attempted to obtain information on the question of how the results vary by examining measurement invariance in terms of the educational backgrounds of teachers both using and not using the method of sub-sampling.

Sub-sampling method

Imbalanced sample sizes in groups may affect the outcome of measurement invariance studies. The difference between the observed model and the estimated model may be disregarded due to the relatively lower weight within the smaller group. Therefore, in case the sample sizes of the groups examined differ greatly, the outcomes of invariance studies may be misleading (Yoon & Lai, 2018).

Chen (2007) found that the power of detecting noninvariance led to a substantial drop when sample sizes in two groups were quite different. Although both of these studies noted potential problems of unbalanced sample sizes in testing factorial invariance, neither included a systematic investigation of unequal sample size conditions that would influence power in detecting violations of invariance (Yoon & Lai, 2018).

Yoon and Lai (2018) suggested that researchers use many random samples from the larger group in testing measurement invariance and report the summary of the results using many random samples. For example, the sub-samples of the larger group may be selected randomly 100 times and each sub-sample selected randomly and the smaller group may be used collectively for measurement invariance analysis. Thus, measurement invariance analysis is conducted 100 times for the different sub-samples of the larger group while using the same sample for the smaller group each time. The fit indices are recorded for each different instance and the average of the fit indices recorded for all 100 instances is calculated. If both the average values and the relevant percentage values for the fit indices are within the range of good fit, the measurement invariance model is supported (Yoon & Lai, 2018).

The R package software was used for creating sub-samples based on the educational background of teachers. The group consisting of teachers with associate degrees, whose size is the smallest (see Table 1) was taken into consideration for the Turkish test. The software output obtained was a file to be used for measurement invariance analysis on Mplus.

Table 2 shows the item averages of student groups created based on the professional experience and education status of teachers for the Turkish test in order to demonstrate the similarity of the averages for the sub-samples acquired using the sub-sampling method with sample averages. Furthermore, the table also features the averages of the sub-samples obtained with the sub-sampling method based on educational background.

Table 2

Item Averages Based on the Educational Background and Professional Experience of Teachers

	Professional Experience			Education Status			Education Status (Sub-sampling Methods)		
	Short	Medium	Extensive	Associate Degree	Bachelor's Degree	Master's Degree	Associate Degree	Bachelor's Degree	Master's Degree
M2	0,51	0,54	0,59	0,55	0,54	0,55	0,55	0,56	0,54
M3	0,31	0,34	0,38	0,38	0,34	0,31	0,38	0,32	0,31
M4	0,70	0,73	0,73	0,77	0,72	0,76	0,77	0,70	0,76
M5	0,47	0,51	0,54	0,61	0,50	0,45	0,61	0,53	0,46
M6	0,70	0,77	0,79	0,84	0,75	0,74	0,84	0,77	0,75
M7	0,65	0,73	0,74	0,79	0,71	0,67	0,79	0,73	0,67
M9	0,46	0,55	0,60	0,61	0,53	0,51	0,61	0,52	0,51

An assessment of Table 2 might lead to the conclusion that the items included in the Turkish test are generally of average difficulty. Additionally, the item averages based on educational background and the averages of the data originating from the educational background sub-samples seem to be close. In other words, the averages of the sub-samples were found to be similar to the average value for the original sample.

Data Collection Process

The open-ended and multiple-choice items included in the ABIDE 2016 assessment were developed by item writers, subject matter experts, measurement and evaluation specialists, and language experts. Then, a pilot scheme was conducted with around 5000 students. The tests were finalized using the item and test statistics at the end of the pilot scheme. Between April and May 2016, the main assessment scheme was put into action in 81 provinces (MEB, 2017). The research data were obtained from the Directorate-General of Measurement, Evaluation, and Examination Services within the MEB.

Data Collection Tools

The study was conducted on the basis of existing data containing the answers given by students to the multiple-choice questions of the Turkish test in the ABIDE 2016 evaluation as well as of teacher questionnaire data. No additional data collection tools were employed besides the ones indicated here. Table 3 shows the number of items per booklet for the ABIDE 2016 assessment.

Table 3

ABIDE 2016 Booklet Types and No. of Items

A Booklet	B Booklet	C Booklet
9 + 9 = 18 items	9 + 9 = 18 items	9 + 9 = 18 items
A1: 18+2 pilot=20 items	B1: 18+2 pilot=20 items	C1: 18+2 pilot=20 items
A2: 18+2 pilot=20 items	B2: 18+2 pilot=20 items	C2: 18+2 pilot=20 items
A3: 18+2 pilot=20 items	B3: 18+2 pilot=20 items	C3: 18+2 pilot=20 items
A4: 18+2 pilot=20 items	B4: 18+2 pilot=20 items	C4: 18+2 pilot=20 items

Source: ABIDE 2016 Report

The present study focuses on the items included in the Turkish test within Booklet A of the ABIDE 2016 assessment. Booklet A consists of nine multiple-choice and nine open-ended questions. The answers given for the open-ended items were scored as incorrect (0), partially correct (1), and correct (2). As for the multiple-choice items, the items were scored as either correct (1) or incorrect (0). The nine multiple-choice items were included within the scope of the study.

Data Analysis

The research data were analyzed in three stages. The first stage involved examining the missing values, outliers, and the number of multicollinearity assumptions. The second stage concerned the establishment of the achievement model for the subject of Turkish and the attempts to verify the said model. As for the final stage, it was about testing the measurement invariance of the models established on the basis of the educational backgrounds and professional experience levels of teachers using the MGCFA method.

Examining Assumptions

Certain assumptions and requirements for the data obtained from the sample must be tested to minimize the problems that may arise prior to the SEM analyses. These can be listed as missing values, outliers, normality, and multicollinearity (Çokluk, Şekercioğlu & Büyüköztürk, 2010).

Missing values

The initial step before continuing with the analyses involved the examination of missing values. There are different approaches for dealing with missing values. The missing data must be completely random for these approaches to be used. 365 students were excluded from the present study because the data were categorical, the sample size was large, and the missing values accounted for less than 5% of the data and were distributed randomly. The missing values within the data used for the study ranged between 0.2% and 1.6%.

Outliers and Normality

The outliers and the assumption of normality were not examined as the data employed in the present study were categorical.

Multicollinearity

For this assumption, the relationships among the items in each factor must be analyzed. A correlation value exceeding 0.90 among the items gives rise to the issue of multicollinearity. A high correlation signifies that the items assess similar properties (Tabachnick & Fidell, 2013). Therefore, the question of whether the correlations were below 0.90 within the tetrachoric correlation matrix was examined. Table 4 shows the tetrachoric correlation matrices.

Table 4
Tetrachoric Correlation Matrix

	T0005	T0006	T0009	T0012	T0013	T0016	T0020
T0005	1						
T0006	0,193	1					
T0009	0,192	0,195	1				
T0012	0,173	0,214	0,192	1			
T0013	0,299	0,259	0,313	0,217	1		
T0016	0,236	0,214	0,242	0,167	0,356	1	
T0020	0,317	0,351	0,305	0,271	0,388	0,311	1

Based on the information given in Table 4, there is no multicollinearity among the items as all correlation values among them are below 0.90. Additionally, the tolerance values and variance inflation factors were examined in consideration of multicollinearity. The assumption is accepted if the tolerance value is greater than 0.1 and the variance inflation factor value is lower than 10 (Tabachnick & Fidell, 2013). Table 5 features the values obtained at the end of the analyses.

Table 5
Tolerance and Variance Inflation Values

Items	VIF	Tolerance
T20160005	1,082	0,925
T20160006	1,079	0,927
T20160009	1,077	0,929
T20160013	1,059	0,945
T20160016	1,133	0,883
T20160017	1,092	0,916
T20160020	1,170	0,855

Table 5 proves that all tolerance values are greater than 0.1 and the variance inflation factor values are lower than 10, indicating the absence of multicollinearity.

All the assumptions were examined and the missing values were excluded from the study. Thus, the dataset was rendered suitable for MGCFA. The stage following these analyses involved the specification of the model. The dataset was subjected to EFA prior to the establishment of the model. Then, the model established was confirmed using CFA and modeled using a path diagram.

Exploratory Factor Analysis

EFA was calculated on the basis of 9 multiple-choice items covered within the scope of the study. It was conducted on the Factor10 software based on the tetrachoric correlation matrices since the data were categorical. The KMO value was calculated as 0.747>0.60 while Bartlett's Test of Sphericity revealed the value of p<0.001 for the Turkish test used for the study. In this regard, one might argue that the dataset is suitable for EFA.

The EFA results revealed that the items are collected under a single factor, which is an expected outcome according to the existing body of literature on achievement tests. However, the factor loadings for items no. T20160017 and T20160001 were calculated to be 0.109 and 0.257, respectively, leading to their exclusion from the study. The explained variance rate was 36.76% after the exclusion of the two items. Seven items in the Turkish test were collected under a single factor named "Achievement in Turkish". Table 6 shows the factor loadings for the items.

Table 6
Item Factor Loadings for the Tests

Items	Factor Loadings
T20160020	0,666
T20160013	0,622
T20160016	0,502
T20160009	0,467
T20160006	0,461
T20160005	0,458
T20160012	0,387

Table 6 shows the factor loadings for the items in the test range between 0.387 and 0.666.

Multi-Group Confirmatory Factor Analysis (MGCFA)

MGCFA provides the researcher with information on both the structural validity of the items and the invariance of this validity among groups (Gregorich, 2006). Therefore, MGCFA was used to examine whether the Achievement in Turkish model satisfies the condition of measurement invariance in terms of the professional experience and educational background of teachers within the scope of the study. The stages of the MGCFA were tested using the Mplus 7 analysis software, using the estimation method of WLSMV. Furthermore, the analyses were conducted based on the tetrachoric correlation matrix generated according to the data. The model created at each level was assessed based on the fit indices of χ^2, RMSEA, CFI, and TLI. In MGCFA, one variable is fixed, and other variable values are left to change. This variable is also called the reference variable (Jöreskog & Sörbom, 1993). Within this context, item no. T20160020 was set to be the reference variable in the study.

While identifying groups based on professional experience within the scope of the study, the amounts of time teachers spent in the profession were categorized as 0-5, 6-15, and 16+ years. Then, durations between 0 and 5 years were identified as "short experience" while those ranging between 6 and 15 years were called "medium-level experience" and periods exceeding 16 years were categorized as "extensive experience". In terms of education status, the groups identified were "Associate Degree", "Bachelor's Degree", and "Master's Degree".

Four hierarchical models or hypotheses are tested in measurement invariance through MGCFA. In each level, the differences between chi-squares and fit indices, which are the prerequisites for advancing into the next step, were examined. Table 7 shows the goodness-of-fit and acceptable fit levels of the fit indices. The difference between the models as far as the fit indices of CFI and TLI are concerned must be between -0.01 and 0.01. The studies by Cheung & Rensvold (2002) and Vandenberg & Lance (2000) indicated that the chi-square difference must be taken into consideration for measurement invariance. In models classified as such, it was also asserted that the use of the changes in the χ^2/Sd value and fit indices would produce more accurate and reliable results.

Results

This section discusses the findings regarding measurement invariance among the student groups formed based on the professional experience levels (short, medium-level, extensive) as well as on education levels (Associate Degree, Bachelor's Degree, Master's Degree) of their teachers both using and not using sub-sampling. The stages of measurement invariance examined through MGCFA were implemented in pairwise groups and the fit indices and the differences between these indices were reported in each invariance level. The order of these levels, as indicated previously, were as follows: configural, metric, scalar, and strict invariance. Table 7 features the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the professional experience of teachers.

Table 7

Fit Indices for the Model Indicating the Success in Turkish Among Student Groups Formed Based on the Professional Experience of Teachers

	Levels of Invariance	χ^2	Sd	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔSd	ΔCFI	ΔTLI
Short-Medium	Configural	67,238	28	0,022	0,987	0,981	-	-	-	-
	Metric	64,682	34	0,017	0,990	0,990	4,819 p=0,567	6	0,003	0,009
	Scalar	80,534	40	0,018	0,987	0,986	16,574 p=0,011	6	0,003	0,004
	Strict	69,843	33	0,019	0,988	0,985	12,828 p=0,076	7	0,001	0,001

Short-Extensive	Configural	56,844	28	0,024	0,984	0,976	-	-	-	-
	Metric	65,462	34	0,022	0,983	0,979	11,063 p=0,086	6	0,001	0,003
	Scalar	74,949	40	0,022	0,981	0,980	10,022 p=0,1237	7	0,002	0,001
	Strict	56,038	33	0,019	0,987	0,984	18,030 p=0,011	7	0,006	0,004
Medium-Extensive	Configural	48,734	28	0,017	0,993	0,989	-	-	-	-
	Metric	54,177	34	0,015	0,993	0,991	8,147 p=0,227	6	0,000	0,002
	Scalar	60,779	40	0,014	0,993	0,992	7,015 p=0,319	7	0,000	0,001
	Strict	53,107	33	0,015	0,993	0,991	9,170 p=0,240	7	0,000	0,001

According to Table 7, the values for short and medium-level experience in the structural equation model were calculated as RMSEA=0.022, CFI=0.987, and TLI=0.981. In the metric invariance model, index values were found to be RMSEA=0.017, CFI=0.990, and TLI=0.990, the chi-square difference (p=0.567) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, while the indices were calculated as RMSEA=0.018, CFI=0.987, and TLI=0.986 and the chi-square difference (p=0.011) was significant, the difference between ΔCFI=0.003 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.019, CFI=0.988, and TLI=0.985, the chi-square difference (p=0.076) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The indices were found to be RMSEA=0.024, CFI=0.984, and TLI=0.976 in the structural equation model for short and extensive experience levels. In the metric invariance model, index values were found to be RMSEA=0.022, CFI=0.983, and TLI=0.979, the chi-square difference (p=0.086) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.022, CFI=0.981, and TLI=0.980, the chi-square difference (p=0.123) was insignificant, and the difference between ΔCFI=0.002 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance mode, while the indices were calculated as RMSEA=0.019, CFI=0.987, and TLI=0.984 and the chi-square difference (p=0.011) was significant, the difference between ΔCFI=0.006 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01).

The indices were found to be RMSEA=0.017, CFI=0.993, and TLI=0.989 in the structural equation model for the instances of medium-level and extensive experience. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.227) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.002 was within the desired range (-0.01 -+0.01). In the scalar invariance model, index values were found to be RMSEA=0.014, CFI=0.993, and TLI=0.992, the chi-square difference (p=0.319) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.240) was insignificant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in the Turkish model among the student groups formed based on the professional experience levels of teachers (i.e., short, medium-level, extensive) fulfilled all the levels of measurement invariance. Table 8 shows the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the educational background of teachers without using the method of sub-sampling.

Table 8

Fit Indices for the Models Indicating the Achievement in Turkish Among Student Groups Formed Based on the Educational Background of Teachers

	Levels of Invariance	$\mathcal{X}^2$	Sd	RMSEA	CFI	TLI	$\Delta\mathcal{X}^2$	ΔSd	ΔCFI	ΔTLI
Associate–Bachelor	Configural	52,318	28	0,015	0,994	0,990	-	-	-	-
	Metric	53,969	34	0,013	0,995	0,994	5,848 p=0,440	6	0,001	0,004
	Scalar	62,620	40	0,013	0,994	0,994	9,306 p=0,157	6	0,001	0,000
	Strict	58,399	33	0,015	0,993	0,992	6,600 p=0,471	7	0,001	0,002
Associate-Master	Configural	27,888	28	0,000	1,000	1,001	-	-	-	-
	Metric	35,863	34	0,015	0,994	0,992	7,759 p=0,256	6	0,006	0,009
	Scalar	46,918	40	0,026	0,976	0,975	11,764 p=0,067	6	0,018	0,017
	Strict	38,928	33	0,027	0,979	0,974	8,341 p=0,303	7	0,003	0,001
Bachelor-Master	Configural	53,508	28	0,016	0,993	0,990	-	-	-	-
	Metric	62,468	34	0,015	0,993	0,991	11,214 p=0,015	6	0,000	0,001
	Scalar	70,774	40	0,015	0,992	0,992	8,821 p=0,183	6	0,001	0,001
	Strict	66,625	33	0,017	0,991	0,989	7,987 p=0,334	7	0,001	0,003

According to Table 8, the indices were calculated as RMSEA=0.015, CFI=0.994, and TLI=0.990 in the structural equation model for the instances of having an associate degree or a bachelor's degree. In the metric invariance model, index values were found to be RMSEA=0.013, CFI=0.995, and TLI=0.994, the chi-square difference (p=0.44) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.013, CFI=0.994, and TLI=0.994, the chi-square difference (p=0.16) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.992, the chi-square difference (p=0.47) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.002 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.000, CFI=1.000, and TLI=1.001 in the structural equation model for the instances of having an associate degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.994, and TLI=0.992, the chi-square difference (p=0.26) was insignificant, and the difference between ΔCFI=0.006 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.026, CFI=0.976, and TLI=0.975, the chi-square difference (p=0.07) was insignificant, and the difference between ΔCFI=0.018 and ΔTLI=0.017 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.979, and TLI=0.974, the chi-square difference (p=0.30) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.016, CFI=0.993, and TLI=0.990 in the structural equation model for the instances of having a bachelor's degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.015, CFI=0.993, and TLI=0.991, the chi-square difference (p=0.015) was significant, and the difference between ΔCFI=0.000 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.015, CFI=0.992, and TLI=0.992, the chi-square difference (p=0.183) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.001 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.017, CFI=0.991, and TLI=0.989, the chi-square difference (p=0.33) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in Turkish model among the student groups formed based on the educational backgrounds of teachers (i.e. associate degree, bachelor's degree, master's degree) fulfilled all the levels of measurement invariance.

The previous Table shows the findings regarding measurement invariance for the models indicating achievement in the subject of Turkish among the student groups formed based on the educational background of teachers without using the method of sub-sampling. For comparative purposes, Table 9 shows the fit indices obtained from the invariance tests regarding the model displaying the achievement in Turkish among the student groups formed based on the educational background of teachers using the method of sub-sampling. As the DIFFTEST command on the Mplus software used to calculate the chi-square difference test for the sample obtained through sub-sampling produced no results, the difference test outcomes were calculated manually, leading to the use of the "≅" sign for indicating the chi-square difference results as they are approximate values.

Table 9

Fit Indices for the Models Indicating the Achievement in Turkish Among Student Groups Formed Based on the Educational Background of Teachers (Sub-Sampling Method)

	Levels of Invariance	$\mathcal{X}^2$	Sd	RMSEA	CFI	TLI	$\Delta\mathcal{X}^2$	ΔSd	ΔCFI	ΔTLI
Associate–Bachelor	Configural	29,975	28	0,015	0,985	0,987	-	-	-	-
	Metric	35,818	34	0,014	0,984	0,990	5,843 p≅0,50	6	0,001	0,003
	Scalar	42,815	40	0,016	0,980	0,986	6,997 p≅0,25	6	0,004	0,004
	Strict	36,450	33	0,018	0,979	0,980	6,365 p≅0,50	7	0,009	0,004
Associate-Master	Configural	28,051	28	0,006	0,997	1,000	-	-	-	-
	Metric	35,978	34	0,013	0,992	0,991	7,927 p≅0,25	6	0,005	0,009
	Scalar	47,015	40	0,026	0,974	0,973	11,037 p≅0,07	6	0,018	0,018
	Strict	38,898	33	0,027	0,978	0,973	8,117 p≅0,30	7	0,004	0,000
Bachelor-Master	Configural	30,178	28	0,017	0,988	0,990	-	-	-	-
	Metric	39,813	34	0,023	0,979	0,978	9,635 p≅0,15	6	0,009	0,012
	Scalar	46,785	40	0,023	0,976	0,978	6,972 p≅0,35	6	0,003	0,000
	Strict	40,383	33	0,027	0,975	0,971	6,402 p≅0,50	7	0,001	0,007

According to Table 9, the indices were calculated as RMSEA=0.015, CFI=0.985, and TLI=0.987 in the structural equation model for the instances of having an associate degree or a bachelor's degree. In the metric invariance model, index values were found to be RMSEA=0.014, CFI=0.984, and TLI=0.990, the chi-square difference (p≅0,50) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.003 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.016, CFI=0.980, and TLI=0.986, the chi-square difference (p≅0,25) was insignificant, and the difference between ΔCFI=0.004 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.018, CFI=0.979, and TLI=0.980, the chi-square difference (p≅0.50) was insignificant, and the difference between ΔCFI=0.009 and ΔTLI=0.004 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.006, CFI=0.997, and TLI=1.000 in the structural equation model for the instances of having an associate degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.013, CFI=0.992, and TLI=0.991, the chi-square difference (p≅0.25) was insignificant, and the difference between ΔCFI=0.005 and ΔTLI=0.009 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.026, CFI=0.974, and TLI=0.973, the chi-square difference (p≅0.07) was insignificant, and the difference between ΔCFI=0.018 and ΔTLI=0.018 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.978, and TLI=0.973, the chi-square difference (p≅0.30) was insignificant, and the difference between ΔCFI=0.004 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01).

The indices were calculated as RMSEA=0.017, CFI=0.988, and TLI=0.990 in the structural equation model for the instances of having a bachelor's degree or a master's degree. In the metric invariance model, index values were found to be RMSEA=0.023, CFI=0.979, and TLI=0.978, the chi-square difference (p≅0.15) was insignificant, and the difference between ΔCFI=0.009 and ΔTLI=0.012 was within the desired range (-0.01 - +0.01). In the scalar invariance model, index values were found to be RMSEA=0.023, CFI=0.976, and TLI=0.978, the chi-square difference (p≅0.35) was insignificant, and the difference between ΔCFI=0.003 and ΔTLI=0.000 was within the desired range (-0.01 - +0.01). In the strict invariance model, index values were found to be RMSEA=0.027, CFI=0.975, and TLI=0.971, the chi-square difference (p≅0.50) was insignificant, and the difference between ΔCFI=0.001 and ΔTLI=0.007 was within the desired range (-0.01 - +0.01).

The RMSEA, CFI, and TLI values indicate that all models display goodness-of-fit while the ΔCFI and ΔTLI values display the necessary conditions for the advancement to the next model. Therefore, the Achievement in the Turkish model among the student groups formed based on the educational backgrounds of teachers (i.e., bachelor's degree and master's degree) fulfilled all the levels of measurement invariance.

In the pairwise comparisons made between student groups created in consideration of the educational backgrounds of teachers, the Achievement in Turkish model fulfilled the conditions for all levels of measurement invariance as was the case in the analysis not using the sub-sampling method. However, the fit indices in the analysis not making use of the sub-sampling method were in a range displaying better fit, signifying that the model is a better fit for the data.

Discussion and Conclusion

The present study examined whether the Turkish test in ABIDE 2016 assessment met measurement invariance among student groups created on the basis of the professional experience and educational backgrounds of teachers. Within this scope, the initial step was to look at the assumptions of MGCFA. After those assumptions were met, the model specified with EFA for both courses was confirmed using CFA. Then, the model was confirmed using CFA for each sub-group under the levels of professional experience and educational background. Finally, each level of measurement invariance was examined in the required order.

The Achievement in the Turkish model satisfied all levels of measurement invariance (i.e., configural, metric, scalar, strict) among the groups of professional experience. This shows that the item factor loadings, item threshold values, and error variances are similar among the student groups created based on the instances of short, medium-level, and extensive professional experience of teachers (0-5 years, 6-15 years, and 16+ years). Within this context, one might argue that the averages, observed variances, and covariances of the scores obtained from students in the Turkish test may be compared and that potential differences in student scores stem from the differences in professional experiences of teachers (i.e., 0-5 years, 6-15 years, 16+ years).

However, the fulfillment of the conditions of measurement invariance in the Achievement in Turkish model among student groups created based on the professional experience of teachers must not be interpreted in a way suggesting that professional experience is the only factor affecting the varying levels of student achievement. According to the results of ABIDE 2016 assessment, student achievement generally increases with the increase in the professional experience of teachers. Similarly, Greenwald, Hedges & Laine (1996) also indicated that teachers with more than five years of professional experience are more productive.

The Achievement in the Turkish model satisfied all levels of measurement invariance (i.e., configural, metric, scalar and strict) among the groups formed based on educational background. This means that the item and factor groups, item factor loadings, item threshold values, and error variances are similar among the student groups created based on the educational backgrounds of teachers (associate degree, bachelor's degree, master's degree). Within this context, the averages, observed variances, and covariances of the scores obtained from students in the Turkish test may be compared and the potential differences in student scores might be attributed to the differences in the education statuses of teachers in terms of having an associate degree, bachelor's degree, or master's degree.

The academic literature on the subject matter reports varying results concerning the positive or negative impact of the educational backgrounds of teachers on student achievement. This is indicated to source from the differentiation in the curricula of master's degree programs (Akyüz, 2006). However, the body of research generally suggests that as the education level of the teacher increases, so does the student achievement. As far as ABIDE 2016 is concerned, the findings state the opposite.

Similarly, to those regarding professional experience, the comparisons concerning the education level of teachers must take other variables into consideration as well. The question of whether the student groups created based on the educational backgrounds of teachers are similar in terms of other variables must be taken into account while interpreting research outcomes. The examination of measurement invariance within the scope of studies making comparisons among groups showing other similarities apart from the property analyzed would provide more information regarding the significance of the comparisons made.

As the number of teachers included in the sample containing those with bachelor's degrees was much higher than other groups, the method of sub-sampling was applied as suggested by Yoon & Lai (2018) by selecting 100 different samples on the R software and their averages were used in the subsequent levels. This allowed the researcher to conduct analyses by equalizing the number of students within the groups of teachers having associate degrees, bachelor's degrees, or master's degrees. Furthermore, the student groups formed based on the education level of teachers were also analyzed without using the method of sub-sampling. Even though both analyses found similar results, the fitness of the model for the data used was reported to be higher in the measurement invariance analysis without using sub-sampling. As the fitness tendency of the data regarding the model increases in the case where sub-sampling is not applied, the method concerned might be used in studies where sub-groups of the sample are distributed unevenly.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

Author Contribution: Süleyman ÜLKÜ-Conceptualization, methodology, analysis, writing & editing, visualization. Burcu ATAR-Conceptualization, methodology, writing-review & editing, supervision.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

References

AERA, APA & NCME. (2014). *The standards for educational and psychological testing.* American Educational Research Association.

Akyüz, G. (2006). Türkiye ve Avrupa Birliği ülkelerinde öğretmen ve sınıf niteliklerinin matematik başarısına etkisinin incelenmesi. *İlköğretim Online. 5*(2), 61-74.

Başusta, N. B. ve Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi. 30*(4), 80-90. https://doi.org/10.24106/kefdergi.2570

Bialosiewicz, S., Murphy, K. & Berry, T. (2013, June). An introduction to measurement invariance testing: resource packet for participants. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

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

Byrne Barbara M. (2004). *Testing for Multigroup Invariance Using AMOS Graphics: A Road Less Traveled, Structural Equation Modeling: A Multidisciplinary Journal, 11*(2), 272-300. https://doi.org/10.1207/s15328007sem1102_8

Chen, F. F. (2007). *Sensitivity of goodness of fit indices to lack of measurement invariance. Structural Equation Modeling, 14,* 464–504. https://doi.org/10.1080/10705510701301834

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

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları.* Pegem Akademi.

Dragan, D. & Topolsek, D. (2014, Haziran). *Introduction to structural equation modeling: review, methodology and practical applications.* The International Conference on Logistics & Sustainable Transport, Celje.

Fraenkel, J. R., & Wallen, N.E. (2006). *How to design and evaluate research in education.* McGraw-Hill.

Gall, J. P., Gall, M. D. & Borg, W. R. (1999). *Applying educational research: A practical guide.* Longman Publishing Group.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). *The effect of school resources on student achievement. Review of Educational Research, 66*(3), 361–396. https://doi.org/10.2307/1170528

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. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Hoyle, R.H. (2012). Model specification in structural equation modeling. In R. H. Hoyle (Ed), *Handbook of Structural Equation Modeling* 126-144. The Guilford Press.

Jöreskog, K. G. & Sörbom, D. (1993). *Lisrel 8: Structural equation modeling with the simplis command language.* Scientific Software International, Inc.

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). *Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. Sociological Methods & Research, 40*(2), 279–310. https://doi.org/10.1177/0049124111405301

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

Millî Eğitim Bakanlığı. (2017). ABIDE 2016 ulusal raporu. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Pyschometrika, 58*(4), 525-543. https://doi.org/10.1007/BF02294825

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). The Guilford Press.

Schmith, N. & Kuljanin, G. (2008). Measurement invariance: review of practice and implication. *Human resources management review, 18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

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

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

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. https://doi.org/10.1037/10222-009

Wu, D. A., 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 Assesment, Research & Evaluation, 12*(3), 1-26. https://doi.org/10.7275/mhqa-cd89

Yoon, M. & Mark H. C. Lai (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(2), 201-213. https://doi.org/10.1080/10705511.2017.1387859