Hacettepe Üniversitesi Eğitim Fakültesi Dergisi

Hacettepe University Journal of Education

e-ISSN: 2536-4758

Multi-Group Confirmatory Factor Analysis via Partial Measurement Invariance Methods

Duygu VARGÖR*, Tuncay ÖĞRETMEN**

Article Information

Received: 20.02.2022

Accepted: 26.09.2023

Online First: 14.10.2023

Published: 31.10.2023

ABSTRACT

In this study, the data obtained from the 6-item survey applied to the participants of the Trends in International Mathematics and Science Study (TIMSS) 2019 Science course were used. The data of the TIMSS 2019 United States of America (USA) sample were compared with[1] the data of countries that have different cultural structures such as Turkey (TUR), Japan (JPN), and Norway (NOR). This study aims to investigate the measurement invariance by constructing structural equation models for the comparisons to be meaningful and reliable. In the study, Exploratory Factor Analysis (EFA) is performed, and Structural Equation Model (SEM) is developed. The measurement invariance was examined by performing confirmatory factor analysis of the measurement tool amongst countries with different cultural structures. The results of data comparison between the USA and these other three countries (TUR, JPN, and NOR) are as follow; between USA and TUR: configural and weak; partially scalar; and partially strict invariance, between USA and JPN: configural and weak; scalar and strict invariance, while between USA and NOR: configural and weak; scalar and strict invariance.

Keywords: Measurement invariance, structural equation modelling (SEM), TIMSS 2019

doi: 10.16986/HUJE.2023.501 Article Type: Research Article

Citation Information: Vargör, D., & Öğretmen, T. (2023). Multi-group confirmatory factor analysis via partial measurement invariance methods. *Hacettepe University Journal of Education, 38*(4), 498-504. doi: 10.16986/HUJE.2023.501

1. INTRODUCTION

For the development of the countries, educating each member of the society is important to ensure social progress. In education, international examination and monitoring methods allow countries to observe their states and make comparisons with other countries. Countries that participate in the PISA and TIMSS exams, which are composed in this context, determine the policies to be applied in the education field based on these data and results. PISA and TIMSS exams question mathematics and science literacy and reading skills of 15-year-old students. The obtained data are important for the evaluation of the characteristics and the quality of the studies in the field of education.

In these studies, conducted between countries and groups, there are differences in culture, socioeconomic level, language, etc. of the groups. The prepared scales are prepared with the assumption that they should measure the same quality in different groups. Thinking that the results obtained in the measurement results are only due to the differences between the individual and the groups causes incomplete interpretation. This difference may be due to the measurement tool.

Confirmatory factor analysis must be performed to present the evidence regarding the characteristics of the measurement tool. In addition, the psychometric characteristics of the measurement tool should be tested with intergroup measurement invariance (Gregorich, 2006).

Measurement equivalence is defined by Mellenbergh (1989), Meredith and Millsap (1992) and Meredith (1993) as follows: Measurement equivalence is the probability of a person having a certain observation score regardless of which group that person is in. As a result, under conditions of measurement equivalence, people with the same score in different groups will have the same observation score. If this condition is not met, it may be difficult to interpret the differences between groups based on actual group differences, whether they are due to bias in the tests or items (Chan, 2000; Somer, 2004; Stark et al., 2006).

Affective and participant-related traits that cannot be directly observed in social sciences, health sciences, and especially in educational studies are called latent traits. Latent traits are tried to be measured by associating them with observable variables. The latent variable is the variable that affects observed variables and reveals the relationship between these observed variables.

* Dr., Ministry of Education, TÜRKİYE. e-mail: vargorr@gmail.com (ORCID: 0000-0002-1039-6630)

** Prof. Dr., Ege University, Faculty of Education, Department of Educational Sciences, Division of Assessment and Evaluation in Education, İzmir- TÜRKİYE. e-mail: tuncay.ogretmen@ege.edu.tr (ORCID: 0000-0001-7783-1409)

Latent variables act as a guide in confirmatory factor analysis and structural equation models. In the model that is constructed with latent variables, the required concordance between the data for the construction of the structure can be presented as proof of the structural validity of the measurement tool (Meredith and Horn, 2001). For instance, examination of the correlation between variables can give an idea about whether the variables measure the same or different structure.

If we examine the theoretical content of the multi-group latent variable common factor structure, the 4-item two-factor structure is shown in two independent groups in Figure 1 as an example. Common factor models represent regression equations for each item for which the common factors are the explanatory variable, and each item is the output. In Figure 1, the ovals represent common factors, rectangles represent items, and triangles represent item means and estimates. Single-headed arrows show regression parameters (λ11, λ12, ..., λ42), common factor averages (κ11, κ12, κ21, κ22) and estimates (τ11, τ12, ..., τ42). The double-headed arrows show the common factor variances (φ11, φ12, φ21, φ22), covariances (ϕ(1,2)1, (ϕ(1,2)2) and item residual variances (θ11, θ12, ..., θ42). In each parameter, the first index of the parameter indicates the common item or common factor, and the second index indicates the group. In the example given in Figure 1, the common two-factor load for two structurally identical groups was examined for four items. Under the assumption of normal distribution, factor means (κ) and variances (φ11) are shown with the parameters given in parenthesises. Each group has varying common factor loads within itself, but the groups are independent of each other.

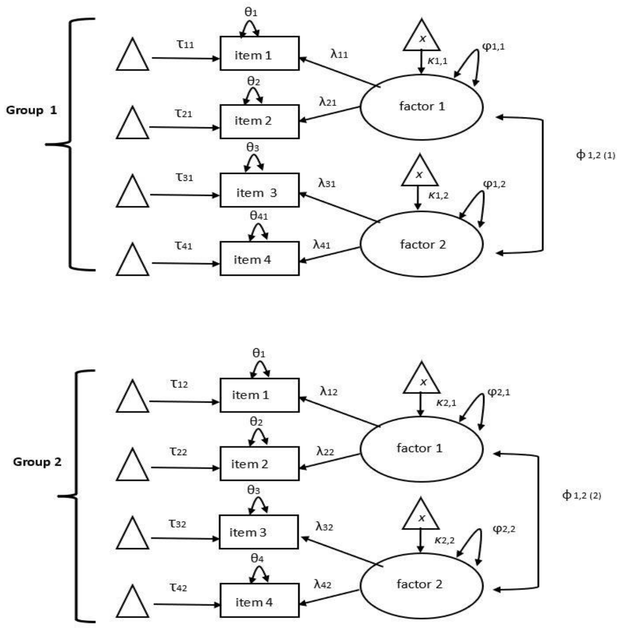

Figure 1. Two-group latent variable common factor

The difference in the measurement results may be due to the measurement tool. A measurement model with the same structure in more than one group means that the factor loadings of the items, their interfactor correlations and the error variances of the scale in question are the same (Jöreskog & Sörbom, 1993; Byrne, 1998). In intergroup comparisons, the invariance of the measurements emerges as a condition for the relevance of this comparison (Bollen, 1989; Cheung & Rensvold, 2000). Primarily, these tests serve to prove whether the obtained measurement results obtained from the measurement tool measure equal structures. Before examining the measurement invariance, it must be examined whether there are equal number of factors in each group in the constructed structural equation models. In structural equation models, the same number of items and factors are preferred. There are four stages in proving measurement invariance (Meredith, 1993). These stages are as follows:

1) Configural invariance: In the first stage of the configural invariance examination, it was determined that the conceptual structures of the compared groups were the same based on the established theory, and the comparison of the basic structure between the groups was evaluated as meaningful and possible. In a word, it is investigated whether the groups have the same factor structure. Factor loadings, regression constants and error variances are released by limiting the number of intergroup factors and loading patterns (Vandenberg & Lance, 2000; Wu et al., 2007). A model is composed in which the number of factors and the model created for the groups are the same in all groups.

2) Weak (Metric) invariance: When metric invariance is investigated, the similarity of the measurement unit of the latent variable is tested. In metric invariance testing, factor loadings are also limited along with the number of factors and loading pattern (Vandenberg & Lance, 2000; Wu et al., 2007). The structural relationships formed for the groups are the same, and the comparison of the structure in terms of variables may be appropriate and meaningful in subgroups.

3) Scalar (Strong) invariance: When scalar or strong invariance is investigated, the number of factors, loading pattern and the regression constant in the constructed regression equation are limited (Vandenberg and Lance, 2000; Wu et al., 2007).

Comparison of latent variables may be appropriate and meaningful since the structural relationship created for the groups and the structural association are the same along the error sources for the subgroups.

4) Strict invariance: When strict invariance is investigated, error variances are limited along with all parameter limitations in measurement invariance (Vandenberg and Lance, 2000; Wu et al., 2007). It is checked whether error variances change between groups.

In the hierarchical four-step invariance stages, for moving to the next step, required fit tests of the previous invariance stage must be provided. Invariance measurement methods that use covariance structure analysis instead of traditional approaches in evaluating differences between groups are used in the comparison of different between-group structures. The most commonly used method in invariance research is Confirmatory Factor Analysis (CFA) (Jöreskog & Sörbom, 1999). CFA tries to ensure the invariance of parameters between groups by making comparisons between models from the least limited model to the most limited model (Horn & McArdle, 1992). It is stated that a study without evidence of invariance in the measurements is considered a weak study (Horn, 1992).

Although the complete proof of invariance has been done in many studies, Byrne and his friends described partial measurement invariance in the CFA structure for the first time (1989). In the study of Gregorich (2006) and the following studies, partial measurement invariance started to be mentioned. According to Gregorich; in a 4-item single-factor study, if the 4th item shows the difference between groups although 3 out of 4-factor loads remain unchanged, it should be defined as partial invariance. Essentially, for some of the items that meet the metric, invariance measurements are used to estimate the related group differences. In accordance with the hierarchical model, it should not be also used in the next invariance step. For instance, items used in the scalar invariance model should not be used in the next one: the strict invariance model. According to Gregorich; in a partially scalar factorial invariance model, only the items that provide scalar factorial invariance criterion contribute to the estimations of group differences in factor means; if observed means are to be compared, only the items that provide scalar factorial invariance criterion should be included in composite measures.

There are many examples about studies on measurement invariance in Turkey. For instance, in Önen's 2009 study, the measurement invariance and partial measurement invariance of the Turkish Form of Epistemological Beliefs Inventory across gender groups were examined. There are also many studies in TUR based on TIMSS and PISA results. For instance, in 2010, Öğretmen and Uzun, conducted measurement invariance tests on the latent variables they found using the Structural Equation Modelling technique in their study with TIMSS TUR data. After examining the fit criterion, it was concluded that the metric invariance condition was met in the model, but the strict invariance condition was not met in the gender groups. In the study by Karakoç, Alatlı, Ayan, Polat-Demir, and Uzun et al (2016) the intercultural measurement invariance in gender differences of TIMSS 2011 mathematics achievement test for 4th graders was examined.

Brancu, Şahin, Guðmundsdóttir, and Çetin et al (2022) evaluated the factor structure and measurement invariance of the Cultural Intelligence Scale (CIS) and analyzed it using data obtained from university students in Romania, Turkey and Iceland. The findings showed a lack of scalar invariance across countries; In other words, it was observed that partial measurement invariance was achieved.

Countries participating in nationwide exams such as TIMSS, PIRLS and PISA gain the opportunity to reach the causes of problems in their education system and to make the necessary changes by using the obtained data. In the PISA-2018 exam, differing from the previous ones, instead of 'Science Achievement', 'Scientific Literacy' was mentioned. The use of the term "Scientific Literacy" instead of "Science" emphasises the significance that PISA attaches to the application of scientific knowledge in the context of real-world situations (MEB, 2019). This new evaluation of 'Scientific Literacy' was updated in the 2015 and 2018 implementations of PISA and became characterized with competence, scientific knowledge, and real-life contexts (content).

It is thought that working with reliable data such as TIMSS, PIRLS and PISA will be important in terms of questioning student skills in the 21st century. According to TIMSS 2019 data, it is examined whether it is meaningful to compare the variables in the model that measures students' affective characteristics towards science for different groups. Research findings will contribute to the literature, and the obtained research findings will make important contributions to the experimental studies carried out for the purpose of change and development in the education system in our country and in other countries.

1.1. The Problem of Study

Score-comparing based structures of internationally and interculturally applied exams such as PISA and TIMSS cause the investigation of their validity. The validity of cross-country comparisons is required to be proven, and this is only possible when the structures of the countries' test scores are measured in the same metric and found to be equal (Meredith and Horn, 2001). Besides, these comparisons are also made on latent traits such as cognitive characteristics, attitudes, and individual differences in psychology and education. For the latent traits of the individuals that are participated in the test to be comparable, the measurement equivalence between groups in terms of their structure must be provided.

Determination of the affective characteristics and qualities of students is merely possible with the evidence obtained from experimental studies. For these pieces of evidence to be used, intergroup comparison of the measured structures alone is not enough. When making comparisons, the investigation of the structures of the groups being compared is also required.

1.2. Purpose of the Study

This study aims to investigate the measurement invariance by constructing structural equation models for the comparisons to be meaningful and reliable. In this study, participant answers to the survey questions that aim to determine the attitudes of the students towards the Science course in the TIMSS 2019 exam were compared. Performing invariance study in the comparison process of groups gives more accurate and reliable results. For these comparisons to be meaningful, measurement invariance should be examined; evaluation of data obtained from both Turkey and other countries with different cultural structures will be enlighten the future studies in terms of the quality of Turkey's science education.

2. METHODOLOGY

2.1. Sample and Data Collection

The data obtained from the TIMSS 2019 USA sample were compared with the data of countries with different cultural structures such as TUR, JPN, and NOR. For comparisons to be meaningful and reliable, structural equation models were created and measurement invariance was examined. The answers given to the questions measuring the affective characteristics about science, which took place in the student survey applied to the 8th grader participants of the TIMSS 2019 international exam by the International Association for the Evaluation of Educational Achievement (IEA), were used as data. In the research, the number of students participating in USA, TUR, JPN and NOR are 8167, 3213, 6802 and 4260, respectively. The countries whose data are used are countries with different cultural structures in which this scale is applied.

In this research, data obtained from a 6-item survey about the TIMSS 2019 Science course were used. This research is descriptive research designed to test the relationship between the variables in the model and success by presenting a model that does not change according to different cultures. For the scaling, 4-point Likert type is used (1-Strongly agree, 2-Somewhat agree, 3-Somewhat disagree, 4-Strongly disagree).

2.3. Data Analysis

Three main analysis methods were used in this study. First, exploratory factor analysis (EFA) was performed. Exploratory factor analysis was done on the 6-item survey of the TIMSS 2019 Science course and ATTITUDE is determined as the latent variable. Before the Exploratory Factor analysis was performed using SPSS 25, KMO (Kaiser-Meyer-Olkin) and Barlett's test of Sphericity were applied.

The KMO value shows that there is a sufficient sample of the item sample in the scale, and the Bartlett Sphericity test shows that the relationship matrix of the test data is factorable. Secondly, in the Structural Equation Model created, the structural relationship between the predictive variables (6 items in the survey) and the latent variable in the factor analysis was examined. With Structural Equation Model (SEM), it is aimed to combine the predictive structural relationship between the variables in the regression model and the latent factor structures in factor analysis with a comprehensive analysis. Third, the measurement invariance of the measurement tool across countries with different cultures was examined with multi-group confirmatory factor analysis (MGCFI).

Gregorich (2006) discussed the "full" version of each factorial form and explained how to analyze partial factorial invariance. In fact, Byrne and colleagues first explicitly defined partial invariance within the CFA framework. Details about the partial factorial invariance method are available in his article.

Data from other countries were compared with the sample of USA, whose native language is English. In this study, differences between the Comparative Fit Index (CFI) were used. The differences between the measurement invariance obtained in the created structural equation model and the CFI were examined respectively. Invariance was examined in the items that are consistent with the data used in the study and has $|\Delta CFI| \leq 0.01$ value and partial invariance was accepted for items that did not fit. (Wu, Li & Zumbo, 2007). In order for ECVI values to be a compatible model; the result of value investigation must be smaller than the ECVI Independent Model values.

3. FINDINGS

The Exploratory Factor analysis was performed using SPSS 25, KMO (Kaiser-Meyer-Olkin) and Barlett's test of Sphericity were applied. Exploratory factor analysis was performed on the 6-item survey of the TIMSS 2019 Science course. The KMO statistic results of USA, TUR, JPN and NOR are 0.917, 0.855, 0.950 and 0.904 respectively. The Bartlett sphericity test results of all groups are all a Sig. (p-value) of 0,000. KMO values show the sample size to be excellent and highly suitable for EFA. Barlett's test of Sphericity shows that the test data come from a multivariate normal distribution and factor analysis can be performed. In the answers of USA, TUR, JPN, and NOR, whose data were analysed, the factor load is found sufficient and factor analysis is found performable. In the analysis, the number of factors with eigenvalues greater than 3 was found to be one factor called ATTITUDE and ATTITUDE is determined as the latent variable. The amount of explained variances of USA, TUR, JPN and NOR datas are % 71, % 55, % 68, % 56 and obtained only one factor.

The data of USA are gradually combined with the data of TUR, JPN and NOR, and the CFI difference value between configural invariance, metric invariance, scalar invariance, and strict invariance is evaluated. The comparison of USA and TUR data is given in Table 1. When the data in Table 1 are examined with the Configural and Weak invariance criteria, ECVI, RMSEA, CFI, GFI and NFI are within the acceptable range (Meredith, 1993). In the given degrees of freedom, the ratio of Chi-Square to the degrees of freedom should be ($\chi^2/sd < 3$). However, in cases where the sample is larger, if other fit indices are provided, the fact that this is being less than 3 loses its importance. Therefore, it is concluded that the measurement model for each variable is compatible for both groups and have the same structure.

However, when the scalar invariance was examined, since $\Delta CFI > 0.01$ and $RMSEA > 0.08$, it was decided that partially scalar invariance was required. Therefore, the regression constant in the regression equation for item 6 is also not limited. As a result of the modification, CFI and RMSEA fit indices are found to be compatible. When the partially strict invariance is examined, it was decided that partially strict invariance should be performed on another item since $\Delta CFI > 0.01$ and $RMSEA > 0.08$. Therefore, the error variances of the 8th item are not limited with the 6th item. As a result of the modification, CFI and RMSEA fit indices were found to be compatible.

Table 1.

Concordance Statistics of the Data from USA and TUR

USA-TUR	sd	χ^2	ECVI	RMSEA	CFI	GFI	NFI	ΔCFI	$\Delta\chi^2$
Configural invariance	18	247,110	0,045	0,064	0,990	0,990	0,990	-	-
Weak invariance	23	262,880	0,054	0,064	0,990	0,990	0,990	0,000	15,770
Scalar invariance	28	472,630	0,170	0,110	0,970	0,990	0,970	0,020	209,750
Partially Scalar invariance (item 6)	27	320,420	0,079	0,073	0,990	0,990	0,990	0,000	57,540
Partially Strict invariance (item 6)	32	492,560	0,120	0,084	0,970	0,980	0,980	0,020	172,140
Partially Strict invariance (item 6&8)	31	390,070	0,091	0,074	0,990	0,990	0,990	0,000	69,650

The data of USA and JPN comparison is given in Table 2. When the data in Table 2 are examined with the Configural, Weak, Scalar and Strict invariance criteria, ECVI, RMSEA, CFI, GFI and NFI are within the acceptable range (Meredith, 1993).

Table 2.

Concordance Statistics of the Data from USA and JPN

USA-JPN	sd	χ^2	ECVI	RMSEA	CFI	GFI	NFI	ΔCFI	$\Delta\chi^2$
Configural invariance	18	247,110	0,042	0,062	0,990	0,980	0,990	-	-
Weak invariance	23	285,820	0,051	0,062	0,990	0,990	0,990	0,000	38,710
Scalar invariance	28	344,420	0,059	0,061	0,990	0,990	0,990	0,000	58,600
Strict invariance	34	471,250	0,088	0,070	0,990	0,980	0,990	0,000	126,830

The data of USA and NOR comparison is given in Table 3. When the data in Table 3 are examined with the Configural and Weak invariance criteria, ECVI, RMSEA, CFI, GFI and NFI are within the acceptable range (Meredith, 1993). In other words, the latent variable ATTITUDE for these two groups is configural and weak invariant. However, when the scalar invariance is examined, it can be observed that ECVI and RMSEA values increased and CFI, GFI and NFI values decreased. Therefore, for item 6, the regression constant in the regression equation is also not limited. As a result of the modification, CFI and RMSEA fit indices are found to be compatible. Therefore, it can be stated that these two data groups are partially scalar invariant in other items except the 6th. When the Partially Strict invariance is examined, the error variance of item 6 is not limited.

Table 3.

Concordance Statistics of the Data from USA and NOR

USA-NOR	sd	χ^2	ECVI	RMSEA	CFI	GFI	NFI	ΔCFI	$\Delta\chi^2$
Configural invariance	18	247,110	0,048	0,067	0,990	0,980	0,990	-	-
Weak invariance	23	304,610	0,059	0,062	0,990	0,970	0,990	0,000	57,500
Scalar invariance	28	552,850	0,110	0,086	0,980	0,970	0,980	0,010	248,240
Partially Scalar invariance (item 6)	27	320,420	0,096	0,081	0,980	0,970	0,980	0,010	15,810
Partially Strict invariance (item 6)	32	515,100	0,100	0,077	0,980	0,970	0,980	0,000	194,680

4. RESULTS, DISCUSSION AND RECOMMENDATIONS

In the first step, all groups have the structural relationship between the predictor variables and the latent variable ATTITUDE in the factor analysis. EFA indicated that the one factor model, ATTIDUTE and it was created SEM Model. Our findings have shown that one latent factor, ATTIDUTE is the same along with the items loaded with their hypothesized construct. In this study, there are some limitations. First of them, the 6-item survey about the TIMSS 2019 Science course used consists of latent factor, but if there were more items and more factors, it would be more comparative study for understanding cross-cultural equivalence and discovering the construct of differences across the cultures. This is a limitation for this research and in future studies it is recommended to be used other scales of TIMSS or PISA exams to explore multidimensional structure. Therefore, we did not examine the relationship between Science course survey and Science exam results across different countries. Future research efforts examining the relationships will be beneficial.

The second limitation is that the TIMSS 2019 Science course was not administered in all countries. Besides, the third limitation is that datas and the answers given obtained from some countries was missing or invalid. As is known, ordinal indicators and missing data are common in applied studies examining latent structures. Unfortunately, the linearity assumption of traditional CFA, which is routinely applied to provide evidence of construct validity of measurement instruments, is violated by ordinal indicators.

(Lei & Shiverdecker, 2020). The countries whose data are used are only four countries of them in this study. So, it was difficult for us to choose the countries in different cultures. It is recommended that measurement invariance will be examined by using the data obtained from different cultures and explored the reason of the missing answers.

In the first stage of the invariance examination, it is determined that all groups are configural invariant. In fact, it means that the number of factors and the model created for the groups are the same in all groups. In the second stage, it is determined that all groups are weak invariant. It is found that the structural relationships formed for the groups are the same, and the comparison of the structure in terms of variables may be appropriate and meaningful in subgroups. The data of USA and JPN are examined and there is no need any modification. The latent variable ATTITUDE for two groups is configural, weak, scalar, and strict invariant. As a result of the examination, comparison of latent variable is appropriate and meaningful since the structural relationship created for the USA and JPN samples and the structural association are the same along the error sources for the subgroups and error variances aren't different in the groups.

After examining the data of USA and NOR, there is need modification for one item. As a result of the modification for item 6, CFI and RMSEA fit indices are found to be compatible. Therefore, two data groups are partially scalar invariant and partially strict invariant in other items except for the 6th item. Comparison of latent variable is appropriate and meaningful since the structural relationship created for the USA and JPN samples and the structural association are the same along the error sources for the subgroups and error variances aren't different in the groups except for the 6th item.

The data of USA and TUR are examined. As a result of the modification for the 6th item, CFI and RMSEA fit indices are found to be compatible. Therefore, two data groups are partially scalar invariant in every item except the 6th. As a result of the modification, CFI and RMSEA fit indices were found to be compatible. Therefore, two data groups are partially strict invariant except for the 6th and 8th items. Based on the comparison between the data obtained from the TIMSS 2019 United States sample and the data obtained from three countries with different cultural structures; TUR, JPN, NOR, the following is concluded: In the Turkey sample, scalar and strict measure invariance is partially achieved, and the answers given to the Item 6, "I look forward to learning Science at school.", and to the Item 8, "I like doing science experiments.", of the Science survey partially provide measurement invariance. The samples of USA and JPN are invariant and the comparisons are reliable. However, after modification of 6th and 8th items, the samples of USA and NOR and the samples of USA and TUR are invariant. If 6th and 8th items are extracted from questionnaire, the comparisons of between countries (USA-NOR and USA-TUR) will be meaningful. The precision and significance of the comparisons cannot be mentioned. We notice that 6th and 8th items are about experiments and the methods of Science lecture. We wonder that it is coincidence or not and it can be investigated in future studies. Therefore, when we consider the reasons that USA-TUR and USA- NOR are scalar invariant and partially strict invariant in other items except for the 6th and 8th items, the relations of the variables not included in the model with the independent variable may have affected the causality relationship. For this reason, the datas will be examined by adding variables like provinces, gender, regions, the social structure etc.

5. REFERENCES

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, *53*(1), 605-634.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review, 110*(2), 203.

Brancu, L., Şahin, F., Guðmundsdóttir, S., & Çetin, F. (2022). Measurement invariance of the Cultural Intelligence Scale across three countries. *International Journal of Intercultural Relations*, *86*, 145-157.

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.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*(2), 169-199.

Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1-27.

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 Suppl 3),* S78.

Horn & Mcardle. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18 (3),117 – 144.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524-544.

Lei, P. W., & Shiverdecker, L. K. (2020). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(4), 584-601.

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*(3), 223-236.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change.

Meredith, W. (1993). MI, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement invariance. *Psychometrika, 57*(2), 289-311.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*(3), 403-424.

Ministry of National Education, (2019). PISA 2018 Turkey Preliminary Report, http://www.meb.gov.tr/meb_iys_dosyalar

Sobel, M. E. (1994). Causal inference in latent variable models.

Somer, O. (2004). Examination of scale equivalence in comparisons between groups: Differentiation of item and test functions. *Turkish Journal of Psychology, 19*(53), 69-82.

Stark, S., Chernyshenko, O. S., Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292- 1306.

Steenkamp, J-BEM, Baumgartner H. (1998). Assessing measurement invariance in cross-national consumer research. J. Consum Res.; 25:78–90.

Uzun, N. B., Gelbal, S., & Öğretmen, T. (2010). TIMSS-R fen başarısı ve duyuşsal özellikler arasındaki ilişkinin modellenmesi ve modelin cinsiyetler bakımından karşılaştırılması. *Kastamonu Üniversitesi Kastamonu Eğitim Dergisi*, *18*(2), 531-544.

Vandenberg, R. J. ve Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69.

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