Eurasian Journal of Educational Research, Issue 66, 2016, 389-406

Examination of the TIMSS 2011 Fourth Grade Mathematics Test in Terms of Cross-Cultural Measurement Invariance[1]

Betul KARAKOC ALATLI[*]
Cansu AYAN[**]
Betul POLAT DEMIR[***]
Gulcin UZUN[****]

Suggested Citation:

Karakoc Alatli, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research*, 66, 389-406 http://dx.doi.org/10.14689/ejer.2016.66.22

Abstract

Problem Statement: Student achievement is considered an indicator of the quality of education, and achievement tests are applied to assess student achievement. International tests are adapted into different languages and cultures with the aim of assessing student achievement on an international level and comparing the achievements of different countries. In our country, a number of tests at the national and international levels are conducted to assess student achievement. One of the tests conducted in our country is called Trends in International Mathematics and Science Study (TIMSS). Countries structure their curricula and education policies based on the results of these studies. However, in order for these comparisons to be meaningful, the constructs measured by the tests should be equivalent. When the relevant literature was examined, it was observed that the number of studies on cross-cultural invariance in Turkey was low and that these studies did not involve TIMSS 2011.

[1]This study was presented at the International Eurasian Educational Research Congress (Istanbul University & EJER, 24-26 April 2014).

[*]Corresponding author, Assist. Prof. Gaziosmanpasa University, Tokat, Turkey, E-mail:betul.alatli@gop.edu.tr
[**]Res.Assist. Ankara University, Ankara, Turkey , E-mail: cayan@ankara.edu.tr
[***]Res.Asist.Dr.OmerHalisdemir University, Niğde, Turkey, E-mail: betul.polat6006@gmail.com
[****]Specialist, MEV College, Ankara,Turkey, E-mail: gulcin.cirakuzun@mevkoleji.k12.tr

Purpose of the Study: The purpose of this study was to examine the measurement invariance of TIMSS 2011 mathematics test in terms of different cultures.

Method: Aiming at examining the intercultural measurement invariance of the TIMSS 2011 mathematics test, this is a survey model that tries to describe an existing situation as it is. The study sample was composed of 1,987 fourth graders from Turkey, England, Japan and the USA. This study was conducted on the data obtained from the TIMSS 2011 mathematics test. Model invariance was examined through multi-group confirmatory factor analysis. LISREL 8.80 for Windows software was used for performance of data analysis.

Findings and Results: The study of measurement invariance was conducted in four steps. It was found that the proposed model was confirmed for all countries, and configural invariance was ensured in the first step, while metric invariance was not ensured in the second step. Therefore, we did not start the scalar invariance or strict invariance analyses. After this step, metric invariance was tested through binary and trilateral combinations in order to determine in which country the invariance was collapsed. It was found that the reason why the metric invariance wasn't ensured was that it was not sourced from only one country.

Conclusions and Recommendations: According to the findings, the invariance across four countries was ensured only in the configural invariance step. Therefore, the items causing the model not to have measurement invariance can be determined, as well as whether the items demonstrated DIF across groups. The items determined to demonstrate DIF can be examined in terms of bias of sources, depending on the expert opinions.

Keywords: Measurement invariance, Multiple-group confirmatory factor analysis, Structural equation modeling

Introduction

Education bears such responsibilities as producing enough quality for a society to maintain its existence and development, preventing the existing values from disappearing, and reconciling the new and old values (Varış, 1998). Education not only ensures social continuity through cultural transmission, but also creates a labor pool that will add novel gains to the cultural heritage and move the society one step forward (Hotaman, 2009). As a result, student achievement is considered as an indicator of the quality of education, and achievement tests are applied to assess student achievement. These tests can be both at the national and international levels. International tests are adapted into different languages and cultures in order to assess student achievement at an international level and compare the achievements of different countries.

In our country, a number of tests at the national and international levels are conducted to assess student achievement. One of the tests conducted in our country is called the Trends in International Mathematics and Science Study-TIMSS, which is organized by the International Association for the Evaluation of Educational Achievement (IEA) whose center is in the Netherlands. TIMSS is a survey focusing on the assessment of student math and science knowledge and skills. It monitors the trends in student achievement in these fields and reveals the differences between national education systems in order to allow education and instruction to be improved. Within the scope of this research, information about education systems, instructional programs and students, teachers and school characteristics are collected, along with data on student performances in mathematics and science (Milli Egitim Bakanligi [MEB], 2015).

Achievement tests and questionnaires involving items aimed at measuring the performance of fourth and eighth graders in math and science took place in TIMSS 2011. In each grade level, there were 14 test booklets. The mathematics tests for fourth graders involved the learning domains of numbers, geometrical shapes, measurement and data display, while for eighth graders, it involved the learning domains of numbers, algebra, geometry, data and probability. The science achievement tests for fourth graders involved the learning domains of life science, physical science and earth sciences, while for eighth graders, it involved the learning domains of biology, chemistry, physics and earth sciences (MEB, 2011). Conducted for the first time in 1995, TIMSS was carried out in 1999, 2003, 2007, and 2011, with the last study in 2015. Table 1 shows the Number of Participating Countries and Turkey's Success Ranking in TIMSS 1999-2015.

Table 1.

Number of Participating Countries and Turkey's Success Ranking in TIMSS 1999-2015

Year	Grade 4			Grade 8		
	Number of Countries	Turkey's Success Ranking		Number of Countries	Turkey's Success Ranking	
		Mathematics	Science		Mathematics	Science
1999	-	-	-	38	31	33
2003	-	-	-	-	-	-
2007	-	-		49	30	31
2011	50	35	36	42	24	21
2015	49	36	35	39	24	21

(MEB, 2003, 2011, 2014a, 2014b, 2016)

Aiming at assessing the achievements of students from different cultures and languages in the disciplines of mathematics and science, in TIMSS, the structures that is measured by the tests is required to be equivalent in order for the comparison to be meaningful. In other words, the basic assumption in intercultural comparisons is that

the tests have measurement invariance (Gierl, 2000). Therefore, the Standards for Educational and Psychological Testing(American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999) and Guidelines on Adapting Tests (Hambleton, 1994; International Test Commission [ITC], 2005) require researchers in intercultural studies to provide evidences of comparability of scores obtained using tools in different languages.

Measurement invariance means that examinees of equal standing with respect to a specific latent structure should on average earn the same test score from items and subscales, irrespective of group membership (AERA, APA, & NCME, 1999). For a test to have measurement invariance, it is required for individuals from different groups whose similar characteristics are measured to have an equal chance of getting a specific score (Millsap, & Kwok, 2004). In other words, a measurement model should have the same construct in different groups, and the tool should have the same items, factor loadings, correlation between factors, and error variance (Jöreskog, & Sörbom, 1993).

Multiple-Group Confirmatory Factor Analysis (MG-CFA) is one of the most preferred methods in testing measurement invariance across groups. MG-CFA involves the simultaneous analysis of a CFA model in more than one group (Brown, 2006). MG-CFA tries to ensure parameter invariance by making comparisons between the least limited models and the most limited models (Horn, & McArdle, 1992, as cited in Uzun-Başusta, 2010). In MG-CFA, the parameters of the measurement model are estimated simultaneously in all groups and are tested as to whether these parameters significantly differ from each other (Jöreskog, & Sörbom, 1993).

Measurement invariance is tested in four steps. These steps are (Meredith, 1993):

1-Configural Invariance: This is the most basic level in measurement invariance. In this first step, whether the groups have the same factor construct is examined. Basic model construct is invariable for the groups. In this model, invariance limitation is not conducted over the estimated parameters. In other words, the groups are permitted to have different parameter values. The configural invariance model has a critical importance because the data will not support the more limiting models if the data do not support the similarity of constant and inconstant parameter pairs across groups (Bollen, 1989).

2-Metric invariance: In this step, whether the different groups respond to the items similarly is examined. It is a limiting model. In this model, factor loadings are limited across groups.

3-Scalar invariance: In this step, whether the obtained regression constant is similar across the groups is examined when the factor score of the groups is zero. In this model, there is constant value/coefficient limitation in addition to the factor loading limitation.

4-Strict invariance: In this last step, whether the error variances differ across the groups is examined. While the strict invariance in the measurement model is tested, error variances are limited along with all parameter limitations.

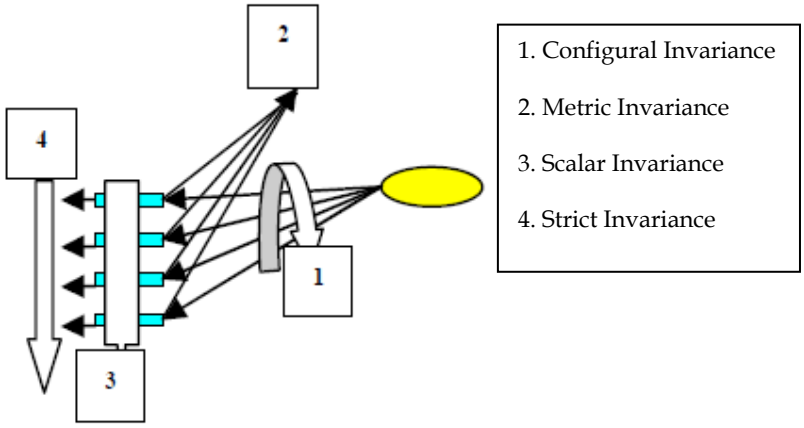

Figure 1. Analysis steps for measurement invariance

Source**:** Başusta, 2010

Vanderberg and Lance (2000) suggested that the evaluation of measurement invariance can be achieved using a systematic approach. This is achieved through a step-by-step process which assesses hypotheses based on their hierarchical order. Every hypothesis is directly related to the specific step in this hierarchal order. Since the steps are in hierarchal order, the structures of the hypothesis are also hierarchal. Therefore, when measurement invariance is not present in one step, there will be no need to evaluate the hypothesis in the next step. Meredith (1993) especially emphasized that full equivalence is a necessary step for a fair and valid comparison. However, full measurement equivalence is generally not used in practice.

Countries structure their curriculums and education policies based on the results of international education studies. However, in order for these comparisons to be meaningful, the constructs measured by the tests should be equivalent. When the relevant literature was examined, it was observed that the number of studies on cross-cultural invariance in Turkey was low (Ogretmen, 2006; Akyildiz, 2009; Asil ve Gelbal, 2012; Asil & Brown, 2015), and none of these studies involved TIMSS 2011. Moreover, it was also determined that measurement invariance was not completely ensured. As a result, it was considered necessary to investigate the cross-cultural measurement invariance of the construct measured by using the TIMMS 2011 mathematics test so that the comparisons would be much more valid and sound. TIMSS is an exam the results of which have an influence on education policy in various countries, and the test also enables countries to compare their levels of education. It is important to determine whether TIMSS shows intercultural

measurement invariance since it is an important cross-cultural exam. There are several advantages of examining the intercultural measurement invariance of TIMSS. The reliability and validity of conclusions derived from TIMSS results will be uncovered. Furthermore, it will enable us to determine how to solve the issues and what the reasons for the problem may be, if any. All of these reasons constitute the necessity to undertake this research.

In this context, the main purpose of this study was to examine the measurement invariance of the TIMSS 2011 mathematics test in terms of different cultures. Within this general purpose, the following questions were examined:

Is there any evidence of TIMSS 2011 in terms of;

a) Configural Invariance

b) Metric Invariance

c) Scalar Invariance

d) Strict Invariance

Method

In this section, information about the research model, population and sample, data collection tool and data analysis are presented.

Research Design

Aiming at examining the intercultural measurement invariance of the TIMSS 2011 mathematics test, this is a survey model since it tries to describe an existing situation as it is.

Research Sample

The target population of TIMSS 2011 consists of all of the fourth and eighth graders in participating countries. The basic sampling model used by TIMSS to obtain a precious and interpretive sample is the two-stage stratified cluster sampling model. The first stage is composed of the selection of schools, while the second stage is composed of selection of classes in those schools.

The population of this study was composed of 50 countries, which participated in TIMSS 2011 at fourth grade level. However, the sample of this study was composed of 1.987 fourth graders from Turkey, England, Japan and the USA, who were selected using purposive sampling methods. The purpose of this selection is that the mother tongues of two countries (England and the USA) are English and the mother tongue of the other two countries (Turkey and Japan) is not English. The element of language, which is one of the most important intercultural differences, has been effective in the selection of countries.

Table 2.

Distribution of Participants by Country

Country	f	%
Turkey	531	26.7
England	250	12.6
Japan	313	15.8
United States of America	893	44.9
Total	1987	100.0

When Table 2 is examined, it can be seen that 531 (26.7%) of the participants are from Turkey, 250 (12.6%) of the participants are from England, 313 (15.8%) from Japan, and 893 (44.9%) from the USA.

Data Collection Tools

This study was conducted usingthe data obtained from the TIMSS 2011 mathematics test. These data were obtained from http://timssandpirls.bc.edu/timss2011/international-database.html. The math questions in TIMSS were limited by numbers, geometrical shapes, measurements and data representation in terms of content. The questions were assessed in three classifications, which are knowledge, application and reasoning in the cognitive domain. TIMMS 2011 Mathematic tests were composed of 14 parallel booklets. The study was carried out using 21 items ona numbered form. Cognitive domain dimensions of items and the number of item in each dimension can be seen in Table 3.

Table 3.

Frequency and Percentage of Fourth Grade Mathematic Items in terms of Cognitive Domain Dimensions

Cognitive Domain	f	%
Knowledge	7	33
Application	6	29
Reasoning	8	38

When Table 3 is examined, it can be seen that 33% of the items were at the knowledge level, 29% were at the application level, and 38% were at the reasoning level.

Data Analysis

LISREL 8.80 for Windows software was used for the data analysis. LISREL was used to create a model and examine invariance across models. Model invariance was examined through multi-group confirmatory factor analysis.

In order to obtain an accurate result from the data, the data set, the data structure and the assumptions of analyses were examined before starting the analysis.

Missing values. First, the missing values were examined since they could lead to great differences in analysis results. The cases having missing values were excluded from research.

Outliers. After missing values, the existence of univariate outliers was examined. It was observed that none of the z scores in any of the cases were within the ± 3 limit. Being a prerequisite for confirmatory factor analysis, multivariate residuals were tested using Mahalanobis Distance. These distances refer to the chi-square distribution whose degree of freedom is the sample size, and they evidence the multivariate outlier observation when the $p < 0.001$ (Kline, 2005; Stevens, 2009).The results showed no multivariate outliers in the data.

Normality. It is difficult to test multivariate normality in Structural Equation Modeling since it requires testing of many linear combinations. In such situations, examination of univariate normality for each observed variable is recommended (Weston &Gore,2006). Skewness and kurtosis values of each variable, and the ratio of mean to the standard deviations (coefficient of variation), were examined to determine the normality of the data. The results demonstrated normal distribution. Graphs about the residuals were examined, and they were decided to be normally distributed. The independence of residuals from each other was examined through Durbin Watson statistics and no test statistic outside the range between 0 and 4 was observed. In this situation, it could be said that the errors were independent of each other(Tabachnick, &Fidell, 2007).

Multicollinearity. The relationship of items to each other and the multicollinearity problems among the items were examined. It was observed that items had low level of relation to each other in each factor. The tolerance values were as expected, while variance inflation factor (VIF) values were below 10 and condition index (CI) values were below 30. These results showed that there was no multicollinearity problem among the items.

Results

The study of measurement invariance was conducted as sequence of testing four steps. The first step is configural invariance, which is the most basic level in measurement invariance, and it examines whether the groups have the same factor construct. The second step is metric invariance in which the different groups respond to the items similarly, and therefore the comparison of different groups' scores can be meaningful. The third step is scalar invariance which expresses that the value of the same subjects has the same value both in latent construct and observed construct. The last step is strict invariance in which the contextual responses given to the factors have invariance.

Meredith (1993) emphasized that strict invariance is required for a fair and valid comparison. However, obtaining the strict invariance is difficult in practice. Therefore, measurement invariance should be expressed gradually. Although there is no language union in this gradation three types of measurement invariance can be determined:

Weak Invariance: for the situation where factor constructs are the same and other parameters are free; Strong Invariance: for the situation where factor constructs and loadings are the same and the error variances are free; Strict Invariance: for the situation where the factor constructs, loadings and error variances are the same (Byrne, Shavelson, &Muthen, 1989).

Does the TIMMS 2011 mathematics test have intercultural measurement invariance?

Configural invariance. In this step, the construct presented in the path diagram in Figure 2 was tested whether to be confirmed or not for the four countries.

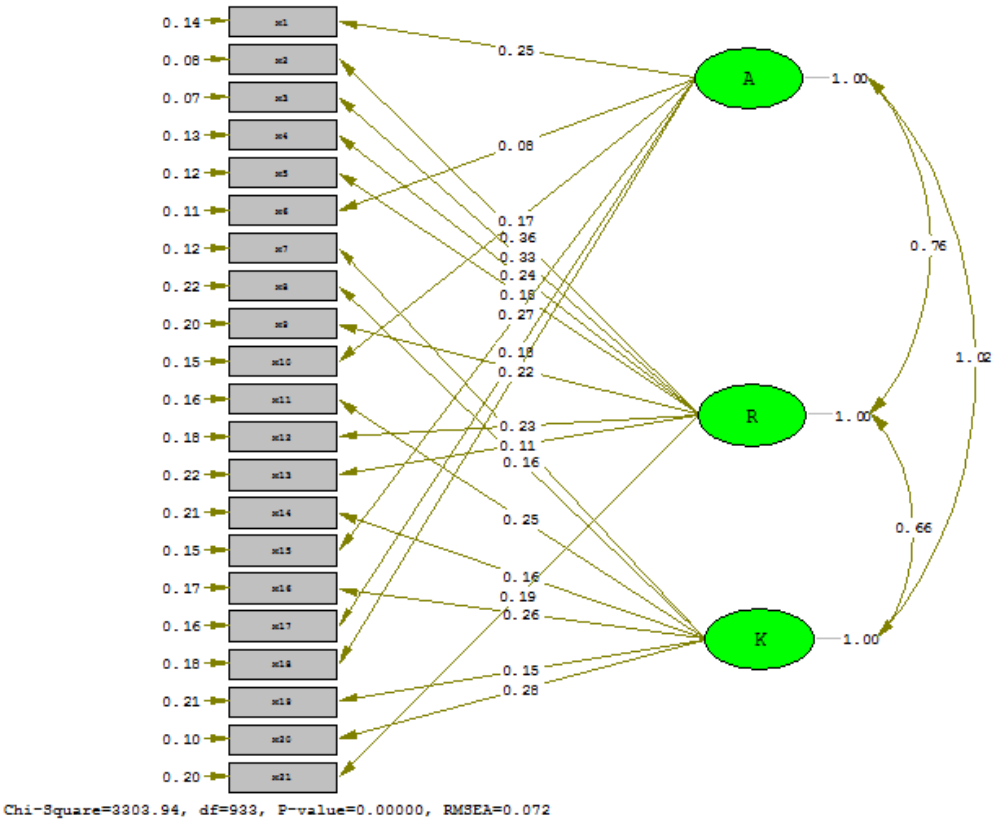

Figure 2. The measurement model of responses given to the mathematics test TIMMS 2011 by students from Turkey, the United States of America, England and Japan

As can be seen in Figure 1, three latent variables were determined related tothe construct tested, which were Knowledge (K), Application (A), and Reasoning (R).

There were 7 indicators of Knowledge latent variable (items 7, 8, 11, 14,16,19,and 20), 6 indicators of Application latent variable (items1, 6, 10, 15, 17,and 18), and 8 indicators of Reasoning latent variable (Items 2,3,4,5,9,12,13,and 21).

Confirmatory factor analysis and configural invariance goodness of fit indexes about the countries are presented in Table 4.

Table 4.

Fit Coefficients of Model about Mathematics Test

Country	χ2 / df	RMSEA	CFI	GFI	RMR	NNFI
Turkey	1.54	0.042	0.97	0.92	0.011	0.96
England	1.22	0.036	0.98	0.89	0.011	0.97
USA	1.84	0.036	0.97	0.95	0.007	0.96
Japan	1.47	0.044	0.96	0.91	0.010	0.95
Configural Invariance	2.44	0.065	0.89	0.86	0.023	0.88

When the Table 4 is examined, it can be seen that the results of confirmatory factor analyses conducted separately for each country showed good fit and the goodness of fit indexes of structural equivalence are at acceptable level ($\chi 2 / df<3$, RMSEA < 0.08, CFI > 0.90, GFI>0.90, RMR< 0.05, NNFI≥0.95). In this area, it can be said that the proposed model was confirmed for all countries and the configural invariance, which is the first step of measurement invariance, was ensured.

Metric invariance. The examination of metric invariance began after the configural invariance was ensured. In the model proposed in this step, factor loadings were fixed for each country, and testing was performed to determine whether the difference between the first situation and the new model was significant. X^2values of the first two steps, degrees of freedom and the differences between them are presented in Table 5.

Table 5.

Fit Coefficients of Metric Invariance Analysis Results by Countries

Step	χ2	df	Δ χ2	Δdf
1. Step	1823.20	748	-	-
2. Step	2206.70	808	383.5	60

As can be seen in Table 5, since $\Delta\chi 2 > 79.08$, the difference between the goodness of fit indexes were significant when the factor loadings were fixed. In other words, metric invariance wasn't ensured. We didn't start the scalar invariance and strict invariance analyses at a step where the metric invariance wasn't ensured since the analysis of measurement invariance is a hierarchical procedure. However, after this

step, metric invariance was tested through binary and trilateral combinations in order to determine in which country the invariance was collapsed.

In order to determine in which country the invariance was collapsed, the metric invariance between three countries was checked after the factor loadings of countries were set free, one by one, respectively. $\Delta\chi2$ and Δdf values with trilateral combinations are presented in Table 6.

Table 6.

Fit Coefficients of Metric Analysis Results by Trilateral Combinations of Countries

Combinations of Countries	$\chi2$	df	$\Delta\chi2$	Δdf
TUR-USA-JPN	2294.02	788	470.82	40
TUR-JPN-ENG	2292.59	788	469.39	40
TUR-USA-ENG	2274.68	788	451.48	40
USA-ENG-JPN	2196.57	788	373.37	40

As can be seen in Table 6, since $\Delta\chi2 > 65.76$, it was observed that metric invariance was not ensured in trilateral combinations of countries. In other words, the reason why the metric invariance was not ensured is not rooted in only one country.

After the metric invariance as not ensured in trilateral combinations of countries, the metric invariance of the four countries was examined in pairs. Fit values, $\Delta\chi2$ and Δdf values of pairs are presented in Table 7.

Table 7.

Fit Indexes of Metric Invariance Analysis Results by Binary Combinations of Countries

Combinations of Countries	$\chi2$	df	$\Delta\chi2$	Δdf
TR-JPN	2211.77	768	388.57	20
TR-USA	2236.67	768	413.47	20
TR-ENG	2201.94	768	378.74	20
USA-ENG	2129.82	768	306.62	20
ENG-JPN	2176.72	768	353.52	20
USA-JPN	2145.98	768	322.78	20

As can be seen in Table 7, since $\Delta\chi2 > 31.41$, it was observed that the metric invariance wasn't ensured in binary combinations. This finding can be interpreted to show that the relationships between characteristics measured and the dimensions of the scale are not similar. In this situation, it can be expressed that the countries did not respond to the items in a similar manner, and making comparison between these scores obtained from these groups is not meaningful.

The configural invariance for the proposed model of the cognitive levels to which the items belonged was ensured. In this step, the differences between the groups can be stated to stem from the measurement tool itself. Therefore, making comparisons across groups may not be accurate. As a result, it can be said that the invariance across countries is weak invariance. This source of this situation is considered to stem from a variety of translation problems and cultural differences. Moreover, it can also be an indicator of Differential Item Functioning (DIF). In the study "Psychometric Properties of Tests for Reading Parts in PIRLS 2001: Turkey and the United States of America (USA)," Ogretmen (2006) determined that the tests did not show any configural invariance among the relevant samples. Their study focused on the intercultural and linguistic invariance of the PISA 2006 student questionnaire. In their study focusing on the intercultural and linguistic invariance of PISA 2006 student questionnaire, Asil and Gelbal (2012) found that some items had differential item functioning across the countries as a result of multiple-group confirmatory factor analysis. As the linguistic and cultural differences increased across countries, it was observed that items demonstrating DIF also increased. The reasons behind the items demonstrating DIF were concluded to be translation problems and cultural differences. In his study focusing on the equivalence of PIRLS 2001 tests across 35 countries, Akyildiz (2009) found that the invariance was ensured at medium level. In a similar study focusing on the examination of TIMMS-R invariance in terms of gender in a Turkish sample, Uzun and Ogretmen (2010) stated that the invariance was ensured except for the metric invariance and that it had a medium level invariance. In the study " The investigation of psychometric properties of the test of progress in international reading literacy (PIRLS) 2001: The model of Turkey-United States of America," Ogretmen (2006) determined that the tests did not show any configural invariance among the relevant samples. As can be seen in similar studies in literature, along with the difficulty of ensuring strict invariance, it was found that metric invariance was mostly ensured, but the equivalence was overruled in scalar invariance, and the medium level invariance was generally ensured. Within the scope of this study, it was observed that only configural invariance was ensured and that it was at a weak level.

Discussion and Conclusion

In this section, the conclusions and recommendations are presented.

Conclusion

In this study, analyses related to the invariance of the model demonstrating the cognitive levels of the TIMMS 2011 mathematics test in Turkey, the USA, England and Japan were conducted. According to the findings, the invariance across four countries was ensured only in the configural invariance step. Metric invariance was tested through binary and trilateral combinations in order to examine in which country the invariance was collapsed in detail, and it was determined that the invariance was not ensured in any combination. Therefore, the invariance across countries was determined to be weak. In this direction, it was concluded that making

comparisons across countries would not be appropriate, the structure of the data should be examined, and troublesome points in terms of culture should be determined.

Recommendations

- Only four countries were selected for this study based on mother tongue. These analyses can involve other countries.
- The items causing the model not to have measurement invariance can be determined, as well as whether the items demonstrated DIF across groups. The items determined to demonstrate DIF can be examined in terms of bias of sources, depending on expert opinions.
- This study took only the language variable into consideration in the cultural comparisons. Other variables may also be included in the research.
- Ensuring the invariance in international examinations such as TIMMS, PISA, and PIRLS is very important for cultural comparisons to be made. Whether there are similar issues in other international examinations can be investigated.

References

Akyildiz, M. (2009). PIRLS 2001 testinin yapi gecerliliginin ulkelerarasi karsilastirilmasi [The comparison of construct validities of the PIRLS 2001 test between countries]. *Yuzuncu Yil Üniversitesi Egitim Fakultesi Dergisi*, 6(1), 18-47.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Asil, M. & Gelbal, S. (2012). PISA ogrenci anketinin kulturler arasi esdegerligi [Cross-cultural equivalence of the PISA student questionnaire].*Egitim ve Bilim*, 37(166), 236-249.

Asil, M., & Brown, G. T. L. (2015). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*. Advance online publication. doi: 10.1080/15305058.2015.1064431

Basusta, B. N. (2010). Olcme esdegerligi. [Measurement invariance].*Egitimde ve Psikolojide Ölcme ve Degerlendirme Dergisi*, 1(2), 58-64.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons, Inc.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

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

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, 25(4), 280-296.

Hotaman, D. (2009). Bazi kisisel degiskenlerin ogrencilerin bagimsiz bir partnerle ve bir grupla calisma aliskanliklariuzerindeki etkisinin arastirilmasi. [The research of the influence of some individual variables on the students' study habits independently, with a partner and in a group]. *Uluslararasi Insan Bilimleri Dergisi,* 6(1).

International Test Commission (2005). *International test commission guidelines for test adaptation*. London: Author.

Joreskog, K., & Sorbom, D. (1993). *Lisrel 8.7.* Chicago: Scientific Software International Inc.

Kline, R.B. (2005), *Principles and practice of structural equation modeling* (2nd Edition ed.). New York: The Guilford Press.

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

Milli Egitim Bakanligi. (2003). TIMSS 1999 Ulusal raporu [TIMSS 1999 National report]. Retrieved October 10, 2016, from http://timss.meb.gov.tr/wp-content/uploads/timss_1999_ulusal_raporu.pdf

Milli Egitim Bakanligi. (2011). TIMSS 2007 Ulusal Raporu [TIMSS 2007 National report].. Retrieved October 10, 2016, from http://timss.meb.gov.tr/?page_id=25

Milli Egitim Bakanligi. (2014a). TIMSS 2011 Ulusal Raporu (4. Siniflar) [TIMSS 2011 National report (4th grades)].Retrieved October 10, 2016, from http://timss.meb.gov.tr/wp-content/uploads/TIMSS-2011-4-Sinif.pdf

Milli Egitim Bakanligi. (2014b). TIMSS 2011 Ulusal Raporu (8. siniflar) [TIMSS 2011 National report (8th grades)]. Retrieved October 10, 2016, from http://timss.meb.gov.tr/wp-content/uploads/TIMSS-2011-8-Sinif.pdf

Milli Egitim Bakanligi. (2016). TIMSS 2015 Ulusal Ön Raporu[TIMSS 2015 National pre-report (4th grades)]. Retrieved December 8, 2016, from http://timss.meb.gov.tr/wp-content/uploads/Timss_2015_ulusal_fen_mat_raporu.pdf

Milli Egitim Bakanligi, (2015). Uluslararasi matematik ve fen egilimleri arastirmasi TIMMS 2011 tanitim kitapcigi [Intoduction booklet of Trends in International Mathematics and Science Study] Retrieved October 10, 2016 from http: egitek.meb.gov.tr/pdf/TIMSS_2011_kitapcigi.pdf

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9(1), 93-115.

Ogretmen, T. (2006). *Uluslararasi okuma becerilerinde gelisim projesi (PIRLS) 2001 testinin psikometrik ozelliklerinin incelenmesi: Turkiye- Amerika Birlesik Devletleri ornegi. [The investigation of psychometric properties of the test of progress in international reading literacy (PIRLS) 2001: The model of Turkey-United States of America].*(Yayinlanmamis Doktora Tezi). Hacettepe Üniversitesi, Ankara.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.

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

Uzun, B., & Ogretmen T. (2010). Fen basarisi ile ilgili bazi degiskenlerin TIMSS-R Turkiye ornekleminde cinsiyete gore olcme degismezliginin degerlendirilmesi [Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey Sample]. *Egitim ve Bilim,* 35(155), 26-35.

Weston, R., & Gore, P. A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist,* 34 (5), 719 – 751.

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.

Varis, F. (1998). *Egitimde program gelistirme [Curriculum development in education].* (7. Basim). Ankara: Alkim Yayincilik.

Zopluoglu, C. (2013). V. Uluslararasi matematik ve fen egilimleri arastirmasi (TIMMS), Turkiye degerlendirmesi: Matematik [Trends in International Mathematics and Science Study (TMSS), Turkey evaluation: mathematics]. *Siyaset, Ekonomi ve Toplum Arastirmalari Vakfi*, 64, 1-14.

TIMSS 2011 Dördüncü Sınıf Matematik Testinin Kültürlerarası Ölçme Değişmezliğinin İncelenmesi

Atıf:

Karakoc Alatli, B., Ayan, C., Polat Demir, B., & Uzun, G. (2016). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in terms of cross-cultural measurement invariance. *Eurasian Journal of Educational Research*, 66, 389-406 http://dx.doi.org/10.14689/ejer.2016.66.22

Özet

Problem Durumu: Eğitim; bir yandan yeni ve eski değerleri bağdaştırarak kültürel aktarımla toplumsal devamlılığı sağlarken; diğer yandan toplumun yaşamasını ve kalkınmasını devam ettirebilecek ölçüde ve nitelikte değer üreterek, kültürel mirasa

yeni kazanımlar ekleyecek insan gücünü yetiştirerek aynı toplumu bir adım ileriye götürmesini sağlamaktadır. Eğitim sonucunda ise öğrenci başarısı, eğitimin niteliğinin bir göstergesi olarak ele alınmakta ve öğrenci başarısının değerlendirilmesinde de başarı testleri uygulanmaktadır. Bu testler ulusal ve uluslararası düzeyde olabilmektedir. Uluslararası düzeyde öğrenci başarılarını değerlendirmek ve farklı ülkelerin başarılarını karşılaştırmak amacıyla hazırlanan uluslararası düzeydeki testler ise farklı dillere ve kültürlere uyarlanmaktadır.

Türkiye'de de öğrenci başarısının değerlendirilmesinde ulusal ve uluslararası düzeyde testler uygulanmaktadır. Uygulanan uluslararası testlerden biri de merkezi Hollanda'da bulunan Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu tarafından düzenlenen Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS)'dır. Öğrencilerin matematik ve fen bilimleri alanlarındaki kazandıkları bilgi ve becerilerini değerlendirmek, eğitimi ve öğretimi geliştirmek amacıyla ülkelerin eğitim sistemleri hakkında karşılaştırmalı veri toplamak TIMSS'in amaçları arasında yer almaktadır. Bu karşılaştırmanın anlamlı olabilmesi için testlerin ölçtüğü yapıların eşdeğer olması yani kullanılan testlerin ölçme değişmezliği/eşdeğerliğinin sağlanmış olması gerekir. Bu bağlamda testlerin, psikometrik bir özellik olarak ölçme değişmezliğine sahip olması, kültürlerarası karşılaştırmalarda, temel bir varsayımdır.

Bir testin ölçme değişmezliğini karşılayabilmesi için, farklı gruplardan gelen fakat benzer yapıları ölçülen bireylerin, belirli bir puanı alma olasılığı eşit olmalıdır. Başka bir deyişle ölçme değişmezliğinin sağlanabilmesi için bir ölçme modelinin birden fazla grupta aynı yapıya sahip olması yani ölçme aracının maddelerinin, faktör yüklerinin, faktörler arası korelasyonlarının ve hata varyanslarının aynı olması gerekir. Ölçme eşdeğerliliği ise dört aşamada test edilir. Bunlar;

1. *Yapısal değişmezlik:* Bu aşamada grupların aynı faktör yapısına sahip olup olmadığı incelenir. Bu modelde kestirilen parametreler üzerinde gruplar arası değişmezlik sınırlandırması yapılmaz yani grupların farklı parametre değerleri almalarına izin verilir.
2. *Metrik değişmezlik:* Bu aşamada, farklı grupların maddelere aynı biçimde cevap verip vermediği incelenir. Bu modelde faktör yükleri gruplar arasında sınırlandırılır.
3. *Skalar değişmezlik:* Bu aşamada özel faktör ortalamalarının yani grupların faktör puanı sıfır olduğunda elde edilen regresyon sabitinin gruplar arasında benzer olup olmadığı incelenir. Bu modelde faktör yükleri sınırlandırmasının yanında sabit değer/katsayı sınırlamasına gidilir.
4. *Tam değişmezlik:* Bu son aşamada hata varyanslarının gruplarda farklılaşıp farklılaşmadığı incelenir. Ölçme modelindeki katı değişmezlik test edilirken bütün parametre sınırlamaları ile birlikte hata varyansları sınırlandırılır

Sonuçları ülke eğitim politikalarına yön vermede ve eğitim programlarının yeniden yapılandırılmasında büyük öneme sahip uluslararası eğitim araştırmalarına dayalı olarak karşılaştırmalar yapabilmek için kullanılan testlerin ölçtüğü yapıların eşdeğer olması gerekmektedir. Literatür incelendiğinde ise kültürlerarası değişmezliğin

incelendiği çalışmaların Türkiye örneklemi için oldukça az olduğu ve bu yapılan çalışmaların TIMSS 2011 uygulamasını kapsamadığı görülmüştür. Bu bağlamda hem testlere dayalı yapılan çıkarımların gerekli ve güvenilir olduğunu belirlemek hem de sorunlar varsa kaynaklarını bulup gidermek açısından TIMSS 2011 uygulamasında yer alantestlerin farklı kültürlerdeki ülkeler arasında ölçme değişmezliğinin sağlanıp sağlanmadığının incelenmesine ihtiyaç duyulmuştur. Bu nedenle TIMSS 2011 Türkiye örnekleminin, anadili İngilizce olan ve olmayan farklı başarı düzeyinde ülkelerle ölçme değişmezliği açısından karşılaştırılması, varsa sorunların belirlenmesi ve daha geçerli güvenilir sonuçlar elde edebilmek ve karşılaştırmalar yapabilmek için yapılabilecek olası çözüm yollarının tartışılması gerekli görülmektedir. Bu amaçla çalışmada, TIMSS 2011 kapsamında yer alan Matematik testinin farklı kültürlerde kültürlerarası ölçme değişmezliği gösterip göstermediği incelenmiştir.

Araştırmanın Amacı: Bu çalışmanın amacı TIMSS 2011 kapsamında yer alan Matematik testinin farklı kültürlere göre ölçme değişmezliğinin incelenmesidir. Bu genel amaç doğrultusunda bu çalışmada şu sorulara yanıt aranmıştır;

TIMSS 2011'in kültürler arası;

a) Yapısal değişmezliğine,

b) Metrik değişmezliğine

c) Skalar değişmezliğine ve

d) Tam değişmezliğine ilişkin kanıt bulunmakta mıdır?

Araştırmanın Yöntemi: TIMSS 2011 kapsamında uygulanan matematik testinde yer alan yapıların kültürlerarası değişmezliğini incelemeyi amaçlayan bu araştırma, var olan bir durumu olduğu şekliyle araştırma söz konusu olduğundan tarama modelindedir. Araştırmanın evrenini TIMSS 2011 uygulamasına 4. Sınıf düzeyinde katılan 50 ülke oluşturmaktadır. Araştırmanın örneklemini ise TIMSS 2011 uygulamasına katılan 50 ülkeden amaçlı örnekleme yöntemi ile belirlenen Türkiye, İngiltere, Japonya ve Amerika Birleşik Devletleri'nden 1987 4. Sınıf öğrencisi oluşturmaktadır. Araştırmaya bu ülkelerin alınmasının amacı iki ülkenin (İngiltere ve Amerika Birleşik Devletleri) anadilinin İngilizce ve diğer iki ülkenin (Türkiye-Japonya) anadilinin İngilizce olmamasıdır. Kültürlerarası en önemli farklılıklardan biri olan dil öğesi, araştırmanın amacı doğrultusunda ülkelerin araştırmaya dahil edilmesinde etkili olmuştur. Araştırma TIMSS 2011 kapsamında uygulanan matematik testi sonuçlarından elde edilen veriler üzerinden yürütülmüştür. Çalışma için gerekli olan veriler http://timssandpirls.bc.edu/timss2011/international-database.html adresinden alınmıştır. TIMMS 2011 Matematik testleri 14 paralel kitapçıktan oluşmaktadır. Araştırma bir numaralı formda yer alan 21 madde ile yürütülmüştür. Maddelerin %33'ü bilme, %29'u uygulama, %38'i ise akıl yürütme alt boyutunda yer almaktadır. Modelin değişmezliği çok gruplu doğrulayıcı faktör analizi ile incelenmiştir. Verilerden doğru bir sonuç çıkartılabilmesi açısından analizlere başlamadan önce veri seti, veri yapısı ve verilerin analizlere ilişkin

varsayımları karşılayıp karşılamadığı incelenmiş, varsayımların karşılandığı sonucuna ulaşılmıştır.

Araştırmanın Bulguları: Bu araştırma kapsamında TIMMS 2011 Matematik maddelerinin bilişsel düzeylerini gösteren modelin Türkiye, Amerika, İngiltere ve Japonya olmak üzere seçilen dört ülkede ölçme değişmezliğinin sağlanıp sağlanmadığına ilişkin analizler yürütülmüştür. Bu anlamda ülkeler arasında hiyerarşik 4 adımdan oluşan değişmezlik kontrolleri yapılmıştır.

1.Yapısal Değişmezlik: İlk adımda kurulan yapının seçilen dört ülke için de doğrulanıp doğrulanmadığı test edilmiştir. Kurulan modelin tüm ülkeler için doğrulandığı ve dolayısı ile değişmezliğin ilk adımı olan yapısal değişmezliğin sağlandığı bulgusuna ulaşılmıştır.

2. Metrik Değişmezlik: Bu adımda kurulan modelde faktör yükleri her ülke için sabitlenmiş ve ilk durum ile yeni modelde elde edilen indeksler arasındaki farkın manidarlığı test edilmiş ve fark manidar bulunmuştur. Yani, metrik değişmezlik sağlanmamaktadır bulgusuna ulaşılmıştır. Değişmezlik analizi hiyerarşik bir yapı gösterdiğinden, metrik değişmezliğin sağlanmadığı adımda analize son verilmiş, skalar değişmezlik ve tam değişmezlik kontrollerine geçilmemiştir. Ancak bu adımdan sonra değişmezliğin hangi ülke ile ilgili olarak bozulduğunu belirleyebilmek adına ülkelerin ikili ve üçlü kombinasyonları arasında metrik değişmezlik incelenmiş ve sağlanmadığı bulgusuna ulaşılmıştır.

Araştırmanın Sonuç ve Önerileri: Araştırma sonucunda, ülkeler arası değişmezliğin zayıf değişmezlik seviyesinde olduğu belirlenmiştir. Bu aşamada yapılan karşılaştırmalarda, gruplar arasındaki farklılıkların ölçme aracından meydana gelebileceği düşünülebilir. Bu doğrultuda, ülkeleri karşılaştırmanın çok uygun olmayacağı, kültürel anlamda sorun çıkarabilecek noktaların tespitinin yapılması gerektiği düşünülmektedir. Bu çerçevede modelin ölçme değişmezliğinin sağlanmamasına neden olan maddeler belirlenerek, gruplar arasında maddelerin DMF (değişen madde fonksiyonu) gösterip göstermediği incelenebilir. DMF gösterdiği tespit edilen maddelerin uzman görüşü alınarak olası yanlılık kaynakları belirlenebilir.

Anahtar Sözcükler: Ölçme eşdeğerliği, Çok gruplu doğrulayıcı faktör analizi, Yapısal eşitlik modeli.