Harran Education Journal,2019
4(1), 1-20
http://dx.doi.org/10.22596/2019.0401.1.20

Assessing Measurement Invariance: Multiple Group Confirmatory Factor Analysis for Differential Item Functioning Detection in Polytomous Measures of Turkish and American Students

Derya Evran[a]

[a]*Harran Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Şanlıurfa,Türkiye*

ABSTRACT

International assessments are often developed in one country and applied in other countries. Assessing the measurement invariance across countries is an important step in determining valid conclusions, comparisons across countries. This paper investigated measurement invariance, across two countries, of selected questions from the Programme for International Student Assessment 2009 student questionnaire. Turkey and United States were compared with the multiple group confirmatory factor analysis for scores on polytomous items to detect differential item functioning (DIF). The results were based on the chi-square goodness of fit test and root mean squared error of approximation, the comparative fit index and the Tucker-Lewis index. The items exhibit DIF, learning strategies, were investigated with Item Response Theory based on the chi-square goodness of fit and t-test.

ARTICLE TYPE

Research

ARTICLE HISTORY

Received 27 February 2019
Accepted 14 March 2019

KEY WORDS

DIF, PISA, MG-CFA, IRT, learning strategies, measurement invariance

Introduction

Large-scale standardized international assessments have become more important in recent decades as a consequence of globalization. In order to compare and evaluate the quality of education and future workforces across countries, many cross-national student assessments have been conducted to compare learning outcomes. One prominent example is the Programme for International Student Assessment (PISA) sponsored by the Organisation for Economic Co-operation and Development (OECD).

Atıf bilgisi: Evran, D. (2019). Assessing measurement invariance: multiple group confirmatory factor analysis for diffirential item functioning detection in polytomous measures of Turkish and American students, *Harran Maarif Dergisi, 4* (1), 1-20. doi: 110.22596/2019.0401.1.20

Sorumlu yazar: Derya Evran, **e-posta:** deryaevran@harran.edu.tr

*Bu makale yazarın yüksek lisans tezinden üretilmiştir.

The reasons for translating or adapting tests in cross-cultural assessments were discussed by Hambleton and Kanjee (1995). Three main advantages of adapting tests across cultures are to improve fairness in assessment by allowing persons to be assessed in the language of their choice, to allow comparison studies of groups at an international and national level, and to reduce costs and save time in developing new tests. However, Hambleton and Kanjee (1995) emphasized that if measurements are not equivalent across different groups, then valid comparisons across these groups cannot be made. Investigation of the differences across cultures, genders, ethnicities, and nationalities is valid only to the degree that assessments provided to the various groups meet the requirements of the measurement invariance (Hambleton and Kanjee, 1995; Wolf, 1998). Measurement invariance or equivalence is defined as the degree to which test scores can be used to make comparable inferences for different examinees (Standards for Educational and Psychological Testing, 1999, p.92). Furthermore, four levels of the measurement invariance were defined: configural invariance, metric invariance, scalar invariance and strict factorial invariance (Meredith, 1993; Van de Vijver and Tanzer, 1997).

Hambleton (2005) identified four cultural/language differences that can affect tests scores: construct equivalence of the test for different cultures, test administration, item formats, and speed effect and emphasized the importance of investigating test and item equivalence. Additionally, Hambleton, Sireci and Patsula (2005) identified three sources of bias in test adaptations: construct bias, method bias, and item bias. Construct bias occurs when a construct is not relevant in all cultures in which the test will be used and when the operational definition varies across cultures. Method bias refers to variation in test administration across cultures, differences in familiarity with the items formats, and differential response styles such as social desirability (Hambleton, 1994). Lastly, item bias refers to faulty translation of items and differential relevance of items across cultures.

Holland and Wainer (1993) defined DIF as different probabilities of answering an item correctly by people who are in different groups but are at the same ability level. In the standards of the American Educational Research Association (1999), DIF was defined as a statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores. In DIF analysis, the focal group is the group of interest, and the reference group is the comparison group of the focal (Holland and Wainer, 1993). The patterns of DIF for polytomous items are more complex because the number of possible response categories is greater than two (Vaughn, 2006).

Measurement invariance under the multiple group CFA model is defined in terms of mathematical equality of the measurement parameters contained in factor loadings, thresholds and residuals that explain the observed variable (Jöreskog, 1971; Meredith, 1993; Steenkamp and Baumgartner, 1998). Muthén and Asparouhov (2002) and Millsap and Yun-Tein (2004) discussed multiple group factor analysis for ordered-categorical variables. Levels of measurement invariance is defined as:

Configural invariance: The number of factors is the same in all groups and the fixed loadings are in the same positions in the matrix of factor loadings (Steenkamp and Baumgartner, 1998). Metric invariance (Weak factorial invariance): The factor loadings are constrained to be equal across groups (Λ 1 = Λ 2) and it implies that the scale intervals are equal across groups (Steenkamp and Baumgartner, 1998). Scalar invariance (Strong factorial invariance): Both factor loadings (Λ 1 = Λ 2) and item intercepts (ν 1 = ν 2) set equal across groups and implies that differences between groups on the observed means are due to the differences between groups on the latent means (Meredith, 1993). Strict factorial invariance: In addition to scalar invariance measurement residual variances are constrained to be equal (Ψ 1 = Ψ 2) across groups (Meredith, 1993).

Additionally, an item response theory (IRT) model specifies the relationship between item scores and a latent ability (Hambleton, Swaminathan, and Rogers, 1991). Two approaches have been proposed to using IRT to detect DIF group differences in IRT item parameters and differences in item response function. When the first approach is used either Wald test or the likelihood ratio test can be used to test equality of the item parameters (Thissen, Steinberg and Wainer, 1988).

The generalized partial credit model (Muraki, 1992) is an IRT model that is for polytomous items. The generalized partial credit model is based on the assumption that for a person with a given level of θ the probability of a score in category *j* rather than category *j-1* is

$$P_j = Prob(Y = j|\theta) = \frac{1}{1 + \exp[-a(\theta - b_j)]} \quad (1)$$

The generalized partial credit model is used to study DIF by allowing "a" (intercept, location) and "bj "(slope) parameters to vary across groups. Measurement invariance of PISA student questionnaire items is important in order to make valid comparisons across countries (OECD, 2012). Despite the obvious importance of investigating measurement invariance for questionnaire items, the approaches to cross-country validation employed in PISA 2009 were fairly limited. The primary purpose of this paper is to investigate measurement invariance of PISA student questionnaire using Multiple Group Confirmatory Factor Analysis (CFA) then further Item Response Theory (IRT) for Differential Item Functioning (DIF) detection. Because the instruments used in international assessments are often originated in developed western countries but applied in developed, newly industrialized, and developing countries and across countries with very different cultures, the constructs assessed may not be equally relevant in all countries, and the meaning of items used to assess the constructs may vary across countries. Two such countries are the United States and Turkey.

For each such question on the PISA 2009 student questionnaire, the following research questions were addressed with the MG-CFA method:

1. Is there any level of measurement invariance between the United States and Turkey?
2. Is there complete measurement invariance between the United States and Turkey?

Further, for each such question on the PISA 2009 student questionnaire that confirm the second question above, the following research questions were addressed with the IRT method:

3. Is there DIF between the United States and Turkey in slope or in location parameters?
4. Is there DIF between the United States and Turkey, item-by-item comparison?

Methods

PISA is an international study, which has a purpose of evaluating educational systems by testing skills and knowledge of 15-year old students in participating countries. PISA was first administered in 2000, to 4500-10000 students in each

participating country every three years. PISA assesses performance in three achievement domains: reading, mathematics and science.

Participants and Data

This paper focused on students who had participated in PISA 2009. All of the PISA datasets are publicly shared by OECD and could be downloaded with the full set of responses from students, school principals, teachers and parents. "These files are accessible to statisticians and professional researchers who would like to undertake their own analysis of the PISA data." (OECD, 2012). The files are available on the OECD webpage and include questionnaires, codebooks, data files in SAS™ and SPSS™ formats, and compendia (http://www.oecd.org/pisa/data/pisa2009database-downloadabledata.htm).

In this paper, PISA 2009 student questionnaire dataset was used and the data for the USA and Turkey was analyzed. Specifically, in the USA, 5233 students from 165 schools in 50 states and 1 district, and in Turkey, 4996 students from 170 schools in 12 geographical regions and 55 provinces were selected. All these participated students responded the same set of the questions in the questionnaire, which were about attitudes towards reading.

Measures

As noted previously, the focus of the present study is on constructs assessed in the student questionnaire. The following lists the specific constructs that were the focus of the study. In total 50 items, each item on a 4-point scale: enjoyment of reading (11 items), learning strategies (13 items) with three factors: memorization strategies, elaboration, and control strategies, teacher-student relations (Five items), disciplinary climate (Five items), teachers' stimulation of reading engagement (Seven items), teachers' use of strategies (Nine items).

DataAnalysis

In this paper multiple group CFA with latent variables and categorical outcomes were used (Millsap and Yun-Tein, 2004; Muthén and Asparouhov, 2002) in order to investigate DIF between countries. For each variable, four models were estimated: the configural invariance, metric invariance, strong factorial invariance, and strict factorial invariance models. The CFA model used to investigate configural invariance in Equations 2 and 3 by using theta parameterization; the variance of ε is set equal to *1* in one group and can be estimated for the second. The further constrains were:

1. For all Y^* set $\nu_g = 0$.
2. Among the factor loadings for each factor, set one loading equal to 1.0. The variable with its factor loading set equal to 1.0 is referred to as the reference variable for the factor.
3. From among the thresholds τ_{g1} to $\tau_{g(j-1)}$ set one equal across the models for the USA and Turkey.

4. For each reference variable, select one additional threshold to be equal across the models for the USA and Turkey.
5. For all variables, set the residual variance to one for one group and estimate the variances in the second group.
6. For one group set the factor means to zero.

A model can be judged adequately fitting either by using the chi-square model fit test or by using model fit indices. The model fit indices were reported: root mean square error of approximation (RMSEA), Comparative Fit Index (CFI) and the Tucker-Lewis Fit Index (TLI). Values at or below .09 for RMSEA, and at or above .90 for CFI and TLI were considered suggestive evidence of good fit. A model was considered adequately fitting if the model comparison test was non-significant or at least two of the three indices meet the less strict criteria for goodness of fit.

The configural invariance model is the least restricted model and the strict factorial invariance model is the most restricted. Adequate fit of the strict factorial invariance model is evidence against claims of DIF. Inadequate fit of the strict factorial invariance model is evidence of DIF.

In IRT context, once we consider two Item Characteristics Curves (ICC)s computed from two different group based on the responses to the same question, if the curves are identifiably different we conclude that the item functions differently for the two groups; hence, we say that the item exhibits DIF. Zumbo (2007) argues in his study "the ICCs can be identifiably different in two common ways" (p. 226). First, the curves can differ only in terms of their threshold (i.e., difficulty) parameter, and hence the curves are displaced by a shift in their location on the theta continuum of variation. Second, ICCs may differ not only on difficulty but also on discrimination (and/or guessing), and hence the curves may be seen to intersect. Within this context, the former represents uniform DIF (i.e., a main effect of group), whereas the latter represents nonuniform DIF (i.e., an interaction of group by ability) However, due to the latent variable modeling approach in the IRT context, a problem arises. This problem is originated from the arbitrariness of the latent trait scale that researcher must set the scale for theta during the calibration. Fortunately, this problem can be overcome using computing algorithms like BILOG-MG, which set the mean of the latent trait distribution at zero.

According to Zumbo (2007) "another issue that arises in IRT DIF is that if the two groups have different ability distributions, then the scales for the groups will be arbitrarily different" (p.227). This is a problem because, in the case of DIF, one wants the two groups on the same scale or

metric. If the groups are not on the same metric, any DIF results will be impossible to interpret. Another study indicated the importance of the groups' ability equality is Angoff (1993) as noted in Muraki (1999); " Differential item functioning (DIF) refers to a test item that displays different statistical properties for different groups after controlling for differences in the abilities of the groups" (p. 217).

Muraki (1999) further discusses that "the standardized DIF measures for slope and item location parameters successfully detect the non-uniform and uniform DIF items as well as recover the means and standard deviations of the latent trait distributions" (p.217). However, there are some common IRT-DIF detection methods described in the literature. These methods include signed area tests (focus on uniform DIF), unsigned area tests (allow for nonuniform DIF), and nested model testing via likelihood ratio test.

Bock and Aitkin (1981) proposed a marginal maximum likelihood (MML) method with the EM algorithm to estimate the IRT parameters. The MML-EM estimation method can avoid the heavy computation of the individual theta values (Muraki, 1999, p.218). MML-EM algorithm used within PARSCALE to estimate the parameters of the multiple-group PCM of this study and the data analyzed based on this model.

For DIF model, it is assumed that different groups have different distributions with mean μ_g and standard deviation σ_g. These distributions are not necessarily normal. The empirical posterior distributions are estimated simultaneously with the estimation of the model parameters (α, β, c). To obtain those parameters, we impose the following constraint for the uniform DIF model for group g:

$$\sum_{j=1}^{J} b_{Rj} = \sum_{j=1}^{J} b_{Fj} \qquad (2)$$

Muraki (1999) states that "this constraint implies the overall difficulty levels of a test or a set of common items given to both reference group R and the focal group F are the same. Therefore, the item difficulty parameters for the focal groups are adjusted. Any overall difference in terms of test difficulty will be assumed to be the difference in ability level for subgroups. The ability level difference among groups can then be estimated by the posterior distributions" (p.222).

For non-uniform DIF, constrain of the slope parameters is applied in a similar fashion. As a result, the DIF measures of item location and slope parameters (β, *and* α) for the focal and reference group for item j are calculated, respectively, by:

$$DIF\ (b_{Fj}) = (b_{Fj^{\wedge}}) - (b_{Rj^{\wedge}})$$
$$DIF\ (a_{Fj}) = (a_{Fj^{\wedge}}) - (a_{Rj^{\wedge}}) \quad (3)$$

Results

PartI: Multiple Group CFA

Evaluation of the models by the chi-square goodness of fit test indicated that all models should be rejected. Model comparison tests for pairs indicated that for each model pair the more restricted model should be rejected in favor of the less restricted model. Table 1 contains the goodness of fit indices for all questions and models. The criterion for adequate fit using RMSEA was 09. The results indicated that for all questions and all models RMSEA was larger than .09 and all models were rejected. Thus, based on RMSEA, none of the models proposed in OECD (2012) adequately fit the data.

Using CFI and TLI, the criterion for adequate fit was greatness from .90, for all questions, except Learning strategies, CFI and TLI indicated that all models fit the data. Further, for all questions except Learning strategies, the change in CFI and TLI was fairly small. These results supported the fit of the strict factorial invariance model and thus that there is no DIF for questions other than Learning strategies. Learning strategies exhibit DIF and the DIF is due either to the thresholds or to the residual variances or to both.

Table 1. Goodness of fit indices for all questions and models

Question	Model	RMSEA	CFI	TLI
Enjoyment of reading	Configural	0.124	0.951	0.939
	Metric	0.124	0.946	0.939
	Strong	0.132	0.926	0.931
	Strict	0.123	0.929	0.940
Learning strategies	Configural	0.096	0.923	0.903
	Metric	0.104	0.901	0.885
	Strong	0.129	0.822*	0.823*
	Strict	0.122	0.827*	0.842*
Teacher student relations	Configural	0.107	0.993	0.986
	Metric	0.150	0.981	0.972
	Strong	0.134	0.975	0.978
	Strict	0.143	0.965	0.975
Disciplinary climate	Configural	0.119	0.989	0.978
	Metric	0.110	0.987	0.981
	Strong	0.117	0.975	0.979
	Strict	0.123	0.967	0.976
Teachers' stimulation of reading engagement	Configural	0.184	0.934	0.902
	Metric	0.167	0.934	0.919
	Strong	0.142	0.935	0.942
	Strict	0.152	0.914	0.933
Teachings' use of strategies	Metric	0.102	0.964	0.958
	Strong	0.166	0.950	0.954
	Strict	0.123	0.925	0.938

Part II: IRT

The focus of the Part II was specifically the questions about learning strategies that are shown in Table 2.

Table 2. Learning strategies questions

1. When I study, I start by figuring out what exactly I need to learn.
2. When I study, I try to relate new information to prior knowledge acquired in other subjects.
3. When I study, I check if I understand what I have read.
4. When I study, I try to figure out which concepts I still haven't really understood.
5. When I study, I try to understand the material better by relating it to my own experiences.
6. When I study, I make sure that I remember the most important points in the text.
7. When I study and I don't understand something, I look for additional information to clarify this.

The base model was no DIF model that is conducted for all 7 items together with using IRT approach for polytomous items. The other 5 models investigate DIF in (i) only location, (ii) only slope, (iii) slope and location, (iv) slope, location and category. The results showed that there is DIF in all items rather than item 1. Models are shown in first column for (i) only slope, (ii) only location, (iii) slope and location, (iv) slope, location and category respectively. According to the results on Table 3, item 2, 3, 4, 5, and 6 have DIF on location, 2, 3, 6, 7 have DIF on slope, 2, 4, 5, have DIF on both in slope and location. By investigating the fit of the measurement invariance models, the results showed suggestive evidence of DIF between the countries; however the evidence about DIF was ambiguous. Thus, further IRT approach was used to detect DIF item by item. Using chi-square model comparison tests between based model and various DIF models, and t-test comparison between item-by-item DIF models the fit is assessed. (The 1's in the model section stands for intercept, slope, and category respectively).

Table 3. DIF models with the items indicating DIF

	Items						
Models	1	2	3	4	5	6	7
1,0,0,0		*	*	*	*	*	
0,1,0,0		*	*			*	*
1,1,0,0		*		*	*		
1,1,1,0		*	*			*	*

The results on Table 3, item 2, 3, 4, 5, and 6 have DIF on location, 2, 3, 6, 7 have DIF on slope, 2, 4, 5, have DIF on both in slope and location. Table 4 shows the results of model comparisons, which shows that all seven items present DIF.

Table 4. T-test comparison between item-by –item DIF models

Items	Slope	Sig.	Location	Sig.
1	1.252		2.119	*
2	-3.059	*	9.371	*
3	1.685		14.281	*
4	7.229		5.029	*
5	-1.108		3.188	*
6	3.216	*	-14.197	*
7	1.861		10.770	*

To conclude, if we freely estimate slope, and put constraints on all the other parameters, the results indicate that items 2, 3, 4, 5 and 6 exhibit DIF. If we freely estimate location and put constrains on all other parameters, the results indicate that items 2, 3, 6 and 7 exhibit DIF. If we freely estimated both slope and location, items 2, 4, and 5 exhibit DIF. In the second procedure, we estimated slope and location freely for 1 item at a time while constraining the parameters on other items between groups. The results for second procedure indicated that items 2, 4, and 6 exhibits DIF in slope and all questions exhibit DIF on location.

Discussion

Investigation of the differences across cultures, genders, ethnicities, and nationalities is valid only to the degree that assessments provided to the various groups meet the requirements of the measurement invariance

(Hambleton and Kanjee, 1995; Wolf, 1998). This paper investigated measurement invariance for several questions on the PISA 2009 student questionnaire by using multiple group CFA. Further, items with DIF were investigated under IRT approach. It is likely that studies have been conducted relating student achievement to questionnaire constructs in PISA data. Even policy changes might happen based on analyses will be proposed. Given this likelihood and noting concern expressed in OECD (2012) about cross-country validity of measures derived from questionnaires, DIF studies are important. To avoid misleading comparisons and conclusions international assessments should be examined carefully.

Learning strategies are the total effort that the students need to process, understand and adopt the information introduced in learning-teaching processes or in their individual preparation (Tay and Harter, 2013). Learning strategy use includes rehearsal (e.g., memorizing material), elaboration (e.g., linking content to other available material), organization (e.g., taking notes, drawing diagrams, or developing concept maps), and monitoring (e.g., monitoring speed and adjusting to time available in an examination) strategies (Pintrich, 2000; Wolters, Pintrich and Karabenick, 2005). With regard to the cognitive engagement, the literature has demonstrated that learning strategy use promotes learning and achievement (Greene, Miller, Crowson, Duke and Akey, 2004; Zimmerman, 2000). Studies have similarly reported that learning strategy use plays a predictive role in achievement in the domain of mathematics (Metallidou and Vlachou, 2007; Wolters and Pintrich, 1998).

The underestimation of the relation between self-reports of learning strategy use and achievement might be also due to the learning strategy use differences between low and high performing students. For example, low performing students might make use of learning strategies in an attempt to do better, high-performing students might tend to respond more differently to the learning strategy questions than do low-performing students in PISA domains. Studies emphasizing the relationship between attitudes and success of students have been conducted in previous PISA cycles (Bybee and McCrae, 2011; Hopfenbeck and Maul, 2011). Moreover, according to OECD, policy adjustments have been planned both in the USA and Turkey (American Educational Research Association, 1999; MEB, 2010). It is likely that studies relating student achievement to questionnaire constructs will be conducted using PISA 2009 data and policy changes will be based on these analyses will be proposed.

Given this likelihood and noting concern expressed in OECD (2012) about cross-country validity of measures derived from questionnaires, additional DIF studies are important.

Conclusions

For the detection of DIF between Turkey and the United States by using multiple group confirmatory factor analysis, the results showed suggestive evidence of DIF between the countries; however the evidence about DIF is ambiguous due to some methodological problems.

First, there is a lack of clear criteria for assessing model fit in confirmatory factor analysis (Browne and Cudeck, 1993; Hu and Bentler, 1999; Kline, 2005). According to Kline (2005), various model fit indices are reported in the structural equation modeling literature, and the minimal sets of fit indices that should be reported are the model chi-square, RMSEA, CFI, and the standardized root mean squared residual (SRMR). Unfortunately SRMR is not available for the approach used in M*plus* when data on polytomous items are analyzed. In addition, there are problems with relying solely on goodness of fit indices. Furthermore, criteria for judging adequacy of fit by using RMSEA, CFI and TLI have not been not firmly established (Hu and Bentler, 1999; Browne and Cudek, 1993). For this study, .06 and .09 were used as criteria for RMSEA, and .95 and .90 criteria were used for CFI and TLI. While RMSEA indicated inadequate fit, CFA and TLI suggested adequate fit in many questions. Also criteria for small, moderate and large changes in goodness of fit indices across invariance models were set forth for use in this study. But these criteria have not been validated. Furthermore, chi-square goodness of fit statistics and chi-square model comparison tests results are dependent on the validity of the graded response model and, in particular, the used of the normal ogive approach to the graded response model. The evaluation of measurement invariance with CFA involves the comparison of relative fits with the chi-square values, but the chi-square value is affected by the validity of the assumption that the graded response model fits the data well. Also evaluation of measurement invariance is affected by sample size. Specifically large sample sizes may lead to rejection of a model even though lack of fit of the model is small and may favor a complex model over a simpler model even when the complex model fits only marginally better than the simpler model (Kline, 2005).

The purpose of this study was to investigate measurement invariance by using factor analysis to determine if there was evidence of DIF between the United States and Turkey in six specific questions from the PISA 2009 student questionnaire. Results for two approaches to selecting well-fitting

models were presented: use of (a) model fit and model comparison tests and (b) model fit indices and comparison of model fit indices. Using both approaches clear evidence of non-invariant measurement models and, therefore, of DIF emerged for the learning strategies question. For the other questions, conclusions about DIF depended on which approach was used. Further, IRT approach was used to clarify whether items on the six questions function similarly for students in the United States and Turkey. The fit of the models were assessed in two ways: (a) by using chi-square model comparison tests between based model and various DIF models and (b) by using t-test comparison between item-by-item DIF models.

References

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

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology.* In P. W. Holland & H. Wainer (Eds.). Hillsdale, NJ: Erlbaum.

Barbara, D., Claus, C. and Manfred, P. (2011). The Role of Content and Context in PISA Interest Scales: A Study of the Embedded Interest Items in the PISA 2006 Science Assessment. *International Journal of Science Education*, *33,* 73-95.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*: 443-459.

Browne, M.W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K.A. & Long, J.S. [Eds.] *Testing structural equation models*. Newbury Park, CA: Sage, 136–162.

Bybee, R. and McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7–26.

Green, B. A., Miller, R. B., Crowson, H. M., Duke, B. L. and Akey, K. L. (2004). Predicting High School Students' Cognitive Engagement and Achievement: Contributions of Classroom Perceptions and Motivation. *Contemporary Educational Psychology, 29*, 462-482.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Measurement methods for the social sciences series, Vol. 2. Fundamentals of item response theory.* Thousand Oaks, CA, US: Sage Publications, Inc.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K. and Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147-157.

Hambleton, R. K. (2005). *Issues, designs, and technical guidelines for adapting tests in multiple languages*. In R. K. Hambleton, P. Merenda, & C. D. Spielberger (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Sireci, S. G. and Patsula, L. (2005). *Statistical methods for identifying flaws in the test adaptation process.* In R. K. Hambleton, P. Merenda, & C. D. Spielberger (Eds.). Hillsdale, NJ: Lawrence Erlbaum.

Hopfenbeck, T. N. and Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*, 95–121.

Hu L. and Bentler P.M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Equation Modeling: *A Multidisciplinary Journal, 6:*1, 1-55.

Holland, P. W. and Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

Kline, R. B. (2005). *Methodology in the social sciences. Principles and practice of structural equation modeling (2nd ed.).* New York, NY, US: Guilford Press.

MEB (2010). *PISA 2009 ulusal ön raporu*. Ankara: Milli Eğitim Bakanlığı-EARGED.

Metallidou, P. and Vlachou, A. (2007). Motivational beliefs, cognitive engagement, and achievement in language and mathematics in elementary school children. *International Journal of Psychology, 42,* 2-15.

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

Millsap, E.R. and Yun-Tein, J. (2004) Assessing Factorial Invariance in Ordered-Categorical Measures, *Multivariate Behavioral Research, 39*:3, 479-515.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple- group partial credit model. *Journal of educational measurement*, *36*(3), 217-232.

Muthén, B. O. and Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (M*plus* Web Note No. 4). Retrieved April 28, 2005, from http://www.statmodel.com/mplus/examples/webnote.html.

OECD (2012). *PISA 2009 technical report.* PISA: OECD Publishing.

Pintrich, P.R. (2000). An Achievement Goal Theory Perspective on Issues in Motivation Terminology, Theory, and Research. *Contemporary Educational Psychology 25*, 92–104.

Steenkamp, J.E.M. and Baumgartner, H. (1998); Assessing Measurement Invariance in Cross-National Consumer Research, *Journal of Consumer Research*, *25*, 1, 1, 78–90.

Tay, L. and Harter, J.K. (2013). Economic and Labor Market Forces Matter for Worker Well‐Being. *Applied Psychology, 5*(2), 193-208.

Thissen, D., Steinberg, L. and Wainer, H. (1988). *Use of item response theory in the study of group differences in trace lines.* In: H. Wainer and H. Braun (Eds). Hillsdale: Lawrence Erlbaum Associates.

Van de Vijver, F. J. R. and Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A bayesian multilevel approach* (Unpublished Doctoral dissertation). Florida State University, Tallahassee, FL.

Wolf, R. M. (1998). Validity issues in international assessments. *International Journal of Educational Research, 29*, 491-501.

Wolters, C.A. and Pintrich, P.R. Instructional Science (1998). Contextual differences in student motivation and self-regulated learning in mathematics, English, and social studies classrooms. *Instructional Science, 26*, 27-46.

Wolters, C. A., Pintrich, P. R. and Karabenick, S. A. (2005). Assessing Academic Self-Regulated Learning. In K. A. Moore & L. H. Lippman (Eds.), *The Search Institute series on developmentally attentive community and society. What do children need to flourish: Conceptualizing and measuring indicators of positive development* (pp. 251-270). New York, NY, US: Springer.

Zimmerman, B.J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology 25*, 82–91.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where is it going. *Language assessment quarterly*, *4*(2), 223-233.

Ölçme Değişmezliğinin Değerlendirilmesi: Türk ve Amerikalı Öğrenciler için Madde İşlev Farklılığının Belirlenmesinde Çoklu Grup Doğrulayıcı Faktör Analizi Kullanımı

Anahtar kelimeler: DIF, PISA, faktör analizi, madde tepki kuramı, öğrenme stratejileri ve ölçme değişmezliği

Amaç

Uluslararası değerlendirmeler genellikle gelişmiş ülkelerce düzenlenmekte ve diğer ülkelerde uygulanmaktadır. Soruların farklı ülkelerde aynı geçerlilikte yorumlanabilmesi için madde işlev farklılığının (DIF) değerlendirilmesi önemli bir konudur. Sonuçların geçerli bir biçimde, her ülkede yorumlanabilmesini belirlemek için madde işlev farklılığı belirli koşullar altında ülkeler arasında karşılaştırmaların yapılabilmesini sağlamaktadır. Uluslararası değerlendirmelerde kullanılan araçlar genellikle gelişmiş batı ülkelerinde hazırlandığı, ancak gelişmiş, yeni sanayileşmiş ve gelişmekte olan ve çok farklı kültürlere sahip ülkelerde uygulandığı için, değerlendirilen yapılar tüm ülkelerde eşit derecede önem taşımayabilir. Büyük ölçekli standartlaştırılmış uluslararası değerlendirmeler küreselleşmenin bir sonucu olarak son yıllarda daha önemli hale gelmiştir. Ülkeler arasında eğitim kalitesini ve gelecekteki iş gücünü karşılaştırmak ve değerlendirmek amacıyla öğrenme sonuçlarını karşılaştırmak için birçok uluslararası öğrenci değerlendirmesi yapılmıştır. Öne çıkan bir örnek, Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından desteklenen Uluslararası Öğrenci Değerlendirme Programı (PISA)'dır. Bu çalışmada Türkiye ve ABD'de, Ekonomik İşbirliği ve Kalkınma Örgütü (OECD)' nün sponsorluğunu yaptığı 2009 yılındaki Uluslararası Öğrenci Değerlendirme Programından öğrenmeye yönelik öğrenci tutumları soruları (PISA student questionnaire items) değerlendirilmiştir. PISA öğrenci anket maddelerinin ölçme değişmezliği, ülkeler arasında geçerli karşılaştırmalar yapmak için önemlidir (OECD, 2012). Anket maddeleri için ölçme değişmezliğinin araştırılmasının önemine rağmen, PISA 2009'da ülkeler arası karşılaştırma yaklaşımları oldukça sınırlı kalmıştır. OECD (2012) teknik raporunda belirtildiği üzere, sadece bilişsel sorulara (matematik, fen, dil bilgisi) madde işlev farklılığı (DIF) uygulanmış, yüksek DIF gösteren maddeler çıkarılmıştır. Bununla birlikte bilişsel sorulara ve öğrenci anketlerine ölçme değişmezliğini (measurement invariance) değerlendiren bir yöntem uygulanmamıştır. Bu çalışmanın temel amacı, PISA öğrenci anketinin Çoklu Grup Doğrulayıcı Faktör Analizi (MG-CFA) ve Madde Tepki Kuramı (IRT) kullanarak ölçme değişmezliğini ve madde işlev farklılığının (DIF) tespitini araştırmaktır.

Yöntem

Madde işlev farklılığının değerlendirilmesi farklı metodlar kullanılarak birçok anket ve ölçek için uygulanmaktadır. Ancak araştırmacılar ölçme değişmezliği ile madde işlev farklılığının ayrı kavramlar olduğunu gözden kaçırabilmektedir. Bu araştırmada ilk basamakta Türk ve Amerikalı öğrencilerin PISA'daki öğrenmeye yönelik tutumuna karşı cevapları Çoklu Grup Doğrulayıcı Faktör Analizi (MG-CFA) kullanılarak, okuma tutumu, öğrenme stratejileri, öğretmen öğrenci ilişkileri, disiplin iklimi, öğretmenlerin okuma katılımını teşvik etmesi ve öğretmenlerin stratejileri kullanması ile ilgili yapılar altında karşılaştırılmış ve ölçme değişmezliği açısından ki-kare uyum testi ve uyumluluk endeksleri ile yorumlanmıştır. Araştırmanın ikinci kısmında ise ölçme değişmezliği basamakları madde işlev farklılığına işaret ettiğinden, belirlenen madde işlev farklılığı (DIF), Madde Tepki Kuramı (IRT) kullanarak yorumlanmıştır.

Bulgular

Sonuçlara göre; OECD (2012) 'de önerilen yapı modelleri, araştırmanın örneklem verilerine yeterince uymamakta ve bazı sorular madde işlev farklılığı (DIF) sergilemektedir. Bununla birlikte, karşılaştırmalı uyum indeksi ve Tucker-Lewis indeksi temel alındığında, öğrenme stratejileri dışındaki sorular madde işlev farklılığı (DIF) göstermektedir. Öğrenme stratejileri dışındaki sorularda, sonuçlar ölçme değişmezliği modelinin uygunluğunu desteklemiş ve Türkiye ve Amerika Birleşik Devletleri'ndeki öğrenciler için bu soruların yorumlanabileceğini göstermiştir. Karşılaştırma sonuçlarına göre, Türk ve Amerikan öğrencileri okuma tutumunda ve öğretmenlerin stratejileri kullanması ile ilgili sorularında büyük, öğretmen öğrenci ilişkileri, disiplin iklimi, öğretmenlerin okuma katılımını teşvik etmesi sorularında küçük farklar göstermiştir.

İlk analizi takip etmesi açısından, ölçme değişmezliği modeline uymayan ve madde işlev farklılığı gösterdiği belirlenen PISA 2009 öğrenme stratejileri soruları, Türkiye ve ABD örneklemleri için madde tepki kuramı (IRT) kullanılarak tekrar karşılaştırılmıştır. Sonuçlar, taban modeli ile DIF modelleri arasındaki grup parametrelerini ki-kare uyum testi ve t-testi ile karşılaştırarak yorumlanmıştır.

Tartışma

Kültürler, cinsiyetler, etnik kökenler ve milliyetler arasında geçerli karşılaştırmalar yapılabilmesi için kullanılan ölçme aracının, karşılaştırılan

gruplar arasında ölçme değişmezliği modelinin gerekliliklerini sağlaması gerekmektedir. (Hambleton ve Kanjee, 1995; Wolf, 1998). Bu çalışmada, PISA 2009 öğrenci anketi üzerinde Çoklu Grup Doğrulayıcı Faktör Analizi kullanarak anket soruları için ölçme değişmezliği incelenmiştir. Sonraki aşamada, DIF'li maddeler IRT yaklaşımı altında incelenmiştir. PISA verilerindeki anket soruları ve öğrenci başarısı ile ilgili çalışmalar yapılması muhtemeldir. Analizlere dayanarak eğitim politikalarında değişikliklerin gerçekleştiği düşünüldüğünde, OECD (2012)'de de ifade edilen ve anketlerden elde edilen sonuçların ülkeler arası geçerliliği ile ilgili kaygılar dikkate alındığında, DIF çalışmaları önemlidir. Yanıltıcı karşılaştırma ve sonuçlardan kaçınmak için uluslararası değerlendirmeler dikkatlice incelenmelidir.

Sonuç

Çoklu Grup Doğrulayıcı Faktör Analizi (MG-CFA) ile belirlenen modeler, ki-kare model uygunluk testi kullanılarak veya model uygunluk indeksleri kullanılarak uygun bir şekilde değerlendirilmektedir. Bu çalışmada rapor edilen model uygunluk indeksleri RMSEA, CFI ve TLI' dır. RMSEA için 0,09 veya daha düşük ve CFI ve TLI için 0,90 veya daha yüksek değerler, uygunluğun kanıtı olarak kabul edilmiştir (Hu and Bentler, 1999; Browne and Cudek, 1993). Model karşılaştırma testi anlamlı değilse veya üç endeksten en az ikisi, uyum için daha yüksek kriterleri karşılıyorsa, uygun bir model olarak kabul edilmiştir. Ölçme değişmezliği değerlendirilmesinde son modelinin yetersiz uyumu, DIF'in kanıtı kabul edilmiştir. Modellerin ki-kare uygunluk testine göre değerlendirilmesi tüm modellerin reddedilmesi gerektiğini göstermiştir. Gruplar arası model karşılaştırma testleri, her model çifti için daha kısıtlı modelin daha az kısıtlı model lehine reddedilmesi gerektiğini göstermiştir. Tablo 1, tüm sorular ve modeller için uygunluk indekslerinin derecelerini içermektedir. Sonuçlar, tüm sorular ve tüm RMSEA modellerinin .09'dan büyük olduğunu ve tüm modellerin reddedildiğini göstermiştir. Bu nedenle, RMSEA'ya dayanarak, OECD'de (2012) önerilen modellerin hiçbiri verilere uygun olmadığı görülmüştür.

CFI ve TLI kriterleri ise Öğrenme stratejileri hariç tüm sorular için .90'dan büyüktür ve modellerin bu soru grubu dışındaki sorular için verilere uygun olduğunu göstermiştir. Ayrıca, öğrenme stratejileri dışındaki tüm sorular için, CFI ve TLI'deki değişim oldukça küçüktür. İlk analiz sonuçlarına göre, öğrenme stratejileri, soru grubunda DIF olduğuna ve eşik değerlere veya varyanslara veya her ikisine de bağlı olduğuna kanaat getirilmiştir. Bu nedenle, ikinci analiz aşamasında DIF bulunan soruları madde madde tespit edebilmek amacıyla IRT yaklaşımı kullanılmıştır. Taban model ve çeşitli DIF modelleri arasındaki ki-kare model karşılaştırma testleri ve DIF modelleri arasındaki t-testi karşılaştırmasının uygunluğu ile sonuçlar değerlendirilmiştir. Tablo 3 ve Tablo 4 madde madde DIF'e sahip olan soruları ve parameter değerlerini göstermektedir.