Eurasian Journal of Educational Research, Issue 58, 2015, 41-60

A Study on Detecting of Differential Item Functioning of PISA 2006 Science Literacy Items in Turkish and American Samples

Nükhet ÇIKRIKÇI DEMİRTAŞLI*

Seher ULUTAŞ**

Suggested Citation:

Çıkrıkçı Demirtaşlı, N. & Uluştaş, S. (2015). A Study on Detecting of Differential Item Functioning of PISA 2006 Science Literacy Items in Turkish and American Samples. *Eurasian Journal of Educational Research*, 58, 41-60. http://dx.doi.org/10.14689/ejer.2015.58.3

Abstract

Problem Statement: Item bias occurs when individuals from different groups (different gender, cultural background, etc.) have different probabilities of responding correctly to a test item despite having the same skill levels. It is important that tests or items do not have bias in order to ensure the accuracy of decisions taken according to test scores. Thus, items should be tested for bias during the process of test development and adaptation. Items used in testing programs, such as the Program for International Student Assessment (PISA) study, whose results are inform educational policies throughout the participating countries, should be reviewed for bias. The study examines whether items of the 2006 PISA science literacy test, applied in Turkey, show bias.

Purpose of the Study: The aim of this study is to analyze the measurement equality of the PISA science literacy test of 2006 in Turkish and American groups in terms of structural invariance and also determined whether the science literacy items show inter-cultural bias.

Methods: The study included data for 15 year-old 757 Turkish and 856 American students. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) was performed to determine whether the PISA science literacy test was equivalent in measurement construct in both groups; multi group confirmatory factor analysis (MCFA) was used to

* Prof. Dr., Department of Measurement and Evaluation, Ankara University, Ankara. rnukhet@yahoo.com

** Dr., Republic of Turkey Ministry of National Education, Chairman of the Board of Education. seherulutas@yahoo.com.tr

identify differences in the factor structure according to cultures. Item bias was detected via the Mantel–Haenszel (MH), Simultaneous Item Bias Test (SIBTEST) and Item Response Theory Likelihood- Ratio Analysis (IRT-LR) procedures.

Findings and Results: : According to the MCFA results PISA 2006 science literacy test for both Turkish and American groups showed equivalent measurement construct. Moreover, the three analyses methods agreed at B and C levels for 15 items in the Turkish sample and 25 items in the American sample in terms of DIF. According to expert opinions, common sources for item bias were: familiarity with item content and differing skill levels between cultures.

Conclusions and Recommendations: The 38 items that showed DIF by each of the three methods were accepted as having DIF. The findings of the present study, possible source of bias in the items will not change the average level of student performance in participating countries. However, it will be beneficial that the review of item content before test administration, in order to reduce the errors items with DIF across different language and cultural groups in international comparative studies.

Keywords: PISA, DIF, Mantel–Haenszel, SIBTEST, IRT-LR

Bias is the presence of some characteristic of an item that results in differential performance for individuals of the same ability in terms of measuring trait but from different ethnic, sex, cultural, or religious groups. In other words, an item biased if equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly. This situation results from some features of items or various situations which are irrelevant with the purposes of the test. Bias is a systematic error affecting the validity of test scores. (Angoff, 1993; Hambleton and Rodgers, 1995; Ellis & Raju, 2003; Reynolds, Livingston & Wilson, 2006).

Items should be tested for potential bias during test construction and adaptation in order to ensure the accuracy of decisions that will be based on the test scores. Methods of determining item bias focus on the validity of test items between particularly different sub-groups (Shepard, Camilli & Williams, 1985). Different methods are used in determining item bias according to classical test theory (CTT) and item response theory (IRT). Within the CTT, many researchers investigated bias by comparing groups via classical statistics such as arithmetic means or item–test correlation. The item bias results obtained by classical methods can vary according to groups, and therefore cannot be generalized to other groups. Thus, researchers have adopted the implicit features model (Embretson & Reise, 2000; Hambleton, Clauser, Mazor & Jones, 1993).In literature on psychometrics, some suggestions were made to use a term other than *bias* for the statistical observation, quite part from its judgmental or interpretive meaning and use, and another term to describe the judgement and evaluation of bias in social sense. Finally the expression *differential*

item functioning came into use, referring to the simple observation that an item displays different statistical properties in different group settings (after controlling for differences in the abilities of the groups) (Angoff, 1993, p.4)

In DIF analysis, the performance of two groups whose skill/competence levels are matched/equivalent is compared for each item. The primary group considers as the focus group and the other is the reference group, which is the basis of the comparison (Donoghue, Holland & Thayer, 1993). Conducting DIF analysis by IRT involves comparison of parameter values estimated from these two groups and the areas between the item characteristic curves estimated from the two groups. In IRT, the item characteristic curve gives a graphical representation of the mathematical function of the correct response pattern and skill measured by items in the test. When the item characteristic curves of an item are not the same for reference and focus groups, the item doesn't measure that proficiency (or ability) similarly in both groups, and hence shows DIF. Item can be interpreted as biased since item characteristic curves will become different when the difference between item parameter values increases (Osterlind, 1983; Camili & Shepard, 1994, Zumbo, 1999; Embretson & Reise, 2000; Baker, 2001).

Bias determination methods based on CTT have advantages and disadvantages relative to IRT (Camilli & Shepard, 1994; Thissen, 2001). Studies generally perform several methods in combination, because previous studies have shown differing outcomes between different tests (Acar, 2008; Ateşok Deveci, 2008; Benito & Ara, 2000; Bertnard & Boiteau, 2003; Doğan & Öğretmen, 2006; Skaggs &Lissitz, 1992; Welkenhuysen-Gybels & Billiet, 2002; Yıldırım, 2006; Bakan Kalaycıoğlu, 2008; Yıldırım & Berberoğlu, 2009). In the present study, the potential for bias within the PISA 2006 science literacy test was investigated by three different methods.

PISA results are taken into consideration by educational policy makers around the world. PISA determines the proficiency of students with 15 year-old in mathematics, science and reading skills at international level. PISA focuses on the competency to use knowledge and skills to overcome difficulties faced in daily life. PISA studies have been conducted at three-year intervals since 2000, and one of mathematics literacy, science literacy and reading skills areas is determined as dominant area in each application period (MEB, 2007; OECD, 2005; MEB, 2010).

Previous studies have reported that the items used in international evaluation studies such as PISA can be subject to bias resulting from translation, adaptation, differences in education programs, etc. (Ercikan, 2002; Ercikan, Mc Creith & Lapointe, 2005; Yıldırım & Berberoğlu, 2009; Le, 2009). The original PISA test was developed in English and translated into the language of participating countries. Thus, language is the most important cultural factor leading to test bias. In this study, whether the items in the PISA 2006 science literacy test conducted in Turkey have any bias suspicion is investigated. The purpose of the research is to determine equality of intercultural (the USA and Turkey) measurement structure of items used in science literacy test in PISA 2006 study as well as the items having bias suspicion from the items used in science test and possible bias reasons by using statistical and judgmental approaches.

Method

The following methods were employed to the research test.

Population and Sampling

Approximately 400,000 students which were included randomly in sampling for representing 20 million students at 15-years old from 57 countries participated to PISA 2006 study. A two-stage stratified sample design was used for the PISA assessment . The first-stage sampling units consisted of schools having 15-year-old students. These schools had selected randomly from seven region in Turkey. Once schools were selected to be in the sample, a complete list of each sampled school's 15-year-old students was prepared. The second-stage sampling units were 15 year-old students within sampled schools. As a result, the Turkish data obtained from 4942, 15 year-old student in 160 schools (OECD, 2005; MEB, 2008).

In this study, the data was used which obtained from 856 American and 657 Turkish students who completed booklet 1 and booklet 5 in the PISA 2006 science literacy test.

Since most of the items in these booklets were released to study by PISA consorcium. the booklets were chosen for this study. These data retrieved from offical PISA web site.

Measures

PISA 2006 science literacy test, which was developed by OECD, as measurement instrument. The PISA test and questionnaires measures higher-order thinking skills such scientific process skills and attitudes towards science. In the PISA test, approximately 40% of items are open ended, 8% are short answers and 52% are multiple-choice. Booklet 1and booklet 5 includes respectively 58 and 60 science literacy items. Of these items, 23 were released; 15 were multiple-choice questions, and 8 were open-ended items (MEB, 2007; MEB, 2010).

Data Analysis

Multiple-choice items were scored as 0–1 and open-ended items were scored as 0–1–2. When using suitable parameter (models for dichotomous items) estimations for items scored with two categories, partial correct and full correct answers were accepted as correct answers and scored by1. in items scored as 1ve 2. Wrong, blank, inaccessible or invalidly marked answers, for example those where more than one option was marked, were coded with 0 as an incorrect response.

Exploratory factor analysis (EFA) was used to determine dimensionality and factor structure of PISA Science literacy test in American and Turkish samples. EFA is generally used to evaluate factor structures or dimensionality of tests in scales and tests (Gierl, 2000; Bolt & Ysseldyke, 2006; Çet, 2006, Yıldırım, 2006). For this purpose, both Principal axis factoring (PAF) and Principal Component factor (PCF) analyse methods were applied on data in order to find a statistical evidence for dimensionality of PISA science literacy tests in each group. The results of PAF showed higher explained total variance for first factor than that provided by PCF

method and also much more items (41items) were loaded under first factor in result of analysis of PAF method. These findings were considered as an evidence for unidimensionality in this study. According to the results, Pisa science literacy test gives a dominant one dimension which has eigenvalues 16,660 for first factor and there was big difference between 1st factor and 2nd factor (eigen value 1,869).

As a second pre-analysis, Confirmatory factor analysis (CFA) was used to prove unidimensionality of the PISA tests and to determine whether the factor structure differs between groups. CFA issued in international studies of factor structures between groups and unidimensionality (Gierl, 2000). Covariance matrices were created in SPSS for CFA via the PRELIS program. The existence of unidimensional structure was controlled for each group and booklet (test) using covariance matrices in the LISREL program (Jöreskog & Sörbom, 1993; Şimşek, 2007). Many studies (Ercikan & Kim 2005; Çet, 2006; Yıldırım, 2008) used multi group confirmatory factor analysis (MCFA) to determine the equivalence of factor structures of tests developed for different cultures. MCFA was used to determine whether the factor structures of PISA Science Literacy test differed with respect to Turkish and American samples.

In this study, one of the DIF analyses was performed as IRT based. Before DIF analysis, PISA data was tested according to IRT basic assumptions; unidimensionality, local independency and model-data fit. In respect to IRT assumptions, data should be one-dimensional structure (Hambleton & Swaminathan, 1985; Gierl, 2000). That's why, the result of PAF method which presented in previous paragraph which was considered as an evidence assumption of unidimensionality in IRT for PISA science literacy test.. In context of PAF results, the eigen values first and second factor were was found respectively, (16.660) and second factor (1.869) and there was small difference between the eigenvalues of the second factor, and third one and the rest (Hambleton & Swaminathan, 1985; Gierl, 2000). Since the PISA data met unidimensionality assumption, another IRT assumption local independency was accepted for the PISA 2006 science literacy test data. (Hambleton, Swaminathan & Rogers, 1991; Osterlind, 1983). In addition to these analyses, PISA data were tested using one-, two- and three-parameter IRT models via the BILOG-MG program in terms of model-data fitting test. The two-parameter model showed best fitting with the data, which had the largest number of items with chi-square value > 0.05.

Mantel–Haenszel (MH) Method. In the MH method and DIF analysis, the performance of two groups was compared by total points (Benito &Ara, 2000; Dorans & Holland, 1992; Donoghue, Holland &Thayer, 1993). The MH D-DIF value, which showed the extent to which the items in tests comprised DIF, was classified according to three categories: A minimal level; B middle level; and C high level. If the item is in category A, MH D-DIF value is zero or less than 1. If the item is in category C, its MH D-DIF value is both bigger than 1.5 and its statistical significance should be more than 1.0. MH D-DIF value between these values is in category B (Dorans & Holland, 1992). During MH analysis, the total scores of the American and Turkish groups were calculated and categorized according to 20% percentile bands. These categories were then used in the EZDIF program developed by Waller (2005).

SIBTEST Method. In the SIBTEST method, items are allocated to two sub-tests: the focus group, comprising items with potential DIF; and the reference group, comprising items not having DIF. For each sub-test point, linear regression is used in order to estimate subtest true scores compared within the scope of "k" focus and reference groups. Estimated true scores are arranged using regression verification techniques (Abbott, 2007; Gierl, Khalig & Boughton, 1999). The following formula gives differences in weighted average between focus and reference groups for subtest item or item clusters examined among k number of subgroups (Abbott, 2007):

$$\hat{\beta}_{UNI} = \sum_{k=0}^{k} p_k d_k$$

Here, P_k is the proportion of focus groups in k number of subgroups; d_k is the difference in adjusted means of item cluster or studies sub-test item for reference and focus groups, respectively, in each k number of sub-groups. If the significance level of $\hat{\beta}_{UNI}$ is positive, DIF is for the reference group; if negative, DIF is for the focus group (Abbott, 2007; Stout, Bolt, Froelich, Habing, Hartz & Roussos, 2003; Zhou, Gierl & Tan, 2005). The value of $\hat{\beta}_{UNI}$ obtained from an item in SIBTEST analysis was classified as follows according to the presence of DIF (Abbott, 2007; Gierl et al., 1999; Gotzmann, Wright & Rodden, 2006): unless there is DIF, the absence hypothesis cannot be rejected and $\left|\hat{\beta}_{UNI}\right|$ is close to zero. When DIF is negligible or at level A, $\left|\hat{\beta}_{UNI}\right| < 0,059$ and $H_0 : \beta = 0$ is rejected. When DIF is at medium level or at level B, $0,059 \le \left|\hat{\beta}_{UNI}\right| < 0,088$ and $H_0 : \beta = 0$ is rejected. When DIF is at significant level or at level C, the value $\left|\hat{\beta}_{UNI}\right| \ge 0,088$ and $H_0 : \beta = 0$ is rejected.

IRT-LR procedures. The IRT-LR method uses a test of statistical significance to compare the differences between two models: compact model (C) and augmented model (A). The purpose of the method is to test whether additional parameters in the augmented method differ from zero. The formula of likelihood rate is as follows:

$$G^2 (df) = 2 \log [\text{Likelihood (A)} / \text{Likelihood (C)}]$$

Here, Likelihood [.] represents the highest likelihood estimation of the parameters of the model; df is the difference between parameter numbers estimated in the compact model and augmented model (Thissen, Steinberg andWainer,1993). In the likelihood proportion statistics for IRT-LR and DIF, the null hypothesis states there is no significant difference between item parameters estimated from two groups. When all parameters are equal that estimated from reference and focus groups, the value of G^2 cannot exceed 3.84 (sd=1, α =0.05for χ^2 distribution). Thus, if the G^2 value exceeds 3.84, the item which considers with DIF (Thissen, 2001). The IRTLRDIF v.2.0b program (Thissen, 2001) was used to determine whether items in the PISA 2006 science literacy test of American and Turkish groups involved DIF according to the IRT-LR method.

Findings and Results

Equivalence of Test Structure. After the equivalence of PISA 2006 science literacy test in Turkish and American samples was detected by the EFA, it was presented by CFA according to chi-square value and goodness of fit statistics for each group and test booklet. These results are given in Table 1.

Table 1

Goodness of Fit Statistics for TURKISH and AMERICAN Samples and Test Booklets

Statistics	TUR		USA		Range for good fit Indices*
	Booklet 1	Booklet 5	Booklet 1	Booklet 5	
χ^2	770.58	931.15	1139.54	907.43	χ^2 df =≤2
df	945	1080	1484	1325	
P	0.99	0.99	1.00	1.00	p> .05
RMSEA	0.00	0.000	0.00	0.000	RMSEA<0.05
AGFI	0.91	0.90	0.91	0.92	AGFI> 0.90
GFI	0.92	0.91	0.91	0.92	GFI> 0.90
CFI	1.00	1.00	1.00	1.00	CFI> 0.90
RMR	0.074	0.076	0.066	0.062	RMR <0.05
NFI	0.75	0.74	0.77	0.82	NFI> 0.90

*(Joreskög and Sörbön,1993; Kelloway, 1998)

As can be seen in Table 1, the value χ^2/ df should be showing unidimensionality of booklets 1 and 5 in Turkish and American groups was non-significant. For the acceptability of a model, the χ^2 value is generally required to be non-significant (Tabachnick & Fidel, 2007). Accordingly, the model was accepted for both groups, so the unidimensional structure existed in both cases. In addition, the RMSEA, AGFI, GFI, RMR and CFI values show that data in both groups are unidimensional.

MCFA was conducted to determine whether the factor structures of tests differed between the Turkish and American Samples. This analysis (Maximum Likelihood-ML) used a covariance matrice since the sample was small and the data was normally distributed. After calculating covariance matrices for each group separately, MCFA was conducted. Three different MCFA models were applied to Booklet 1 data. Model A was applied to determine the equivalence of factor loads, inter-factors correlations and error variances. The results showed that chi-square significance level was not appropriate for three dimensional model. Model B was applied, assuming that correlation between factors and error variances were invariable by releasing the values about *factor loads* to determine which dimension produced the difference between groups. Model B worked better, since the difference was significant at .05 level when comparing Model A and B. However, the model again gave poor fit values to the data. Model C was applied, in which inter-factor correlations were kept held constant by allowing *error variances* in addition to *factor load values* to differ in both groups. Significance tests of the difference between Model

B and C at 0.05 level showed that Model C performed better. Also, considering p likelihood value and goodness of fit values, the model has acceptable goodness of fit, as shown in Tables 2 and 3.

Table 2
Results for Booklet 1 MCFA

Booklet 1	χ^2	df	p	RMSEA
Model A (factor values, inter-factors and error variances are equal)	4800.80	1560	0.00	0.071
Model B (equivalence of inter-factors with error variances)	4750.71	1530	0.00	0.072
Model C (invariance of inter-factor correlation)	3200.76	1490	0.00	0.053

Table 3
Model Comparison for Booklet 1

Model Comparison	χ^2	df
Model A - Model B	50.09	30
Model B - Model C	1549.95*	40

*p<.01

According to these results, the *factor load values and error variances* are different in both groups but factor structures in both groups are the same in terms of inter-factor correlations.

Considering MCFA Booklet 5 and equivalence of *factor values, inter-factors correlation and error variances* of both groups, Table 4 shows that chi-square significance level and other fit values fit the data well. Consequently, all three models showed that the factor structure of booklet 5 data was the same between the Turkish and American samples. According to these results, it was concluded that there was generally a unidimensional structure and that factor structures were equivalent between cultures.

Table 4
Results for Multiple Group Confirmatory Factor Analysis

Statistics	TUR -ABD		Range for Good fit indices*
	Booklet 1	Booklet 5	
χ^2	3200	1343.52	$5 \leq \chi^2/sd \leq 2$
df	1490	1806	
RMSEA	0.053	0.00	0.08<RMSEA < 0.05-
GFI	0.77	-	GFI> 0.90
CFI	-	1.00	CFI> 0.90
RMR	0.06		0.08<RMR < 0.05
NFI	-	0.81	NFI> 0.90

* (Joreskög and Sörbön,1993; Kelloway, 1998)

Differential Item Functioning. Analysis the above tables show the results of the analysis conducted with three methods in order to determine whether the items of PISA 2006 scientific literacy test show intercultural DIF in USA and Turkish groups. All DIF statistics were interpreted at a significance level of α= 0.05. The items showing DIF at B and C levels were taken as DIF, because DIF at levels B and C determine potential bias of the test more sensitively than level A (Gierl et al., 1999; Gotzmann, 2002; Çet, 2006; Gotzmann et al., 2006).

DIF Analysis by Mantel–Haenszel Method. MH analyses are given in Table 5.

Table 5

DIF Analysis by MH Method According to Turkish and American Groups

		Items Numbers/DIF Level	
		B	C
Booklet 1	In favor of Turkish group	5, 13, 19, 33, 36, 38, 44, 45, 49, 53	6, 12, 41
	In favor of American group	4, 10, 15,18, 28, 29, 39, 40, 42, 57, 58	17, 20, 37, 56
Booklet 5	In favor of Turkish group	2, 11, 14, 20, 25, 43, 54, 55, 58, 59	12, 23, 29, 33, 39, 48
	In favor of American group	6, 22, 35, 36, 40,41, 49, 57	3, 5, 8, 16, 46, 47, 52, 60

Examining Table 5, it is seen that 21 of 58 items in booklet 1 show DIF at level B, i.e., at medium level, and 7 items show DIF at level C, i.e., at high level. Of the items showing DIF at B and C levels, 13 were found to be in favor of Turkish students while 15 items were in favor of American students. Table 5 shows that 18 of 60 items in booklet 5 show DIF at level B, while 14 items show DIF at level C. The results indicate that 16 items showing DIF at levels B and C were in favor of Turkish students, while 16 items were in favor of American students.

DIF Analysis by SIBTEST Method. Table 6 shows results for items showing DIF as a result of SIBTEST analysis.

Table 6

Results of DIF Analysis Turkish and American Groups Via SIBTEST Method

		Items Numbers/DIF Level	
		B	C
Booklet 1	In favor of Turkish group	11, 16, 23, 24, 44, 47, 56,	5, 12, 13, 19, 33, 36, 38, 41, 49, 53, 57, 58
	In favor of American group	8, 10, 27,	4, 6, 15, 17, 18, 20, 28, 29, 37, 39, 45,
Booklet 5	In favor of Turkish group	9, 15, 27, 53, 58	2, 11, 12, 14, 20, 23, 25, 29, 33, 39, 43, 48, 54, 55
	In favor of American group	1, 49	3, 5, 6, 8,16, 22, 35, 36, 38, 40, 41, 46, 47, 52, 57, 59, 60

Table 6 shows that 10 of 58 items in booklet 1 showed DIF at level B, while 23 items showed DIF at level C. Of the items showing DIF at levels B and C, 19 were in favor of Turkish students while 14 items were in favor of American students. Of the 60 items in booklet 5, seven involve DIF at level B while 31 items involve DIF at level C. Among the items showing DIF at levels B and C, 19 were in favor of Turkish students while another19 worked in favor of American students.

DIF Analysis by IRT-LR Method. As a result of performing MH and SIBTEST methods, three items that did not show DIF in either of the booklets were taken as "anchor" items, comprising: items 1, 2 and 3 in booklet 1 and the items 1, 4 and 7 in booklet 5. The results of IRT-LR analysis of items including DIF are given in Table 7.

Table 7

Results of DIF Analysis by IRT-LR Method according to Turkish and American Groups

		Items Numbers/DIF Level	
		B	C
Booklet 1	In favor of Turkish group	6, 12, 41,	-
	In favor of American group	4, 10, 15, 17, 18, 21, 28, 29, 37, 39, 40, 42, 49, 56, 57	20
Booklet 5	In favor of Turkish group	2, 12, 14, 29, 33, 48, 57, 58,	-
	In favor of American group	5, 6, 8, 16, 20, 23, 35, 36, 38, 39, 40, 41, 46, 47, 52, 59, 60	-

As sees in Table 7, 18 of 58 items in booklet 1 showed DIF at level B while 1 item showed DIF at level C. Three items showing DIF at levels B and C were in favor of Turkish students while 16 items were in favor of American students. Of the 60 items in booklet 5, it was found that 25 showed DIF at level B and no item showed DIF at level C. Eight items showing DIF at level B were in favor of Turkish students while 17 items were in favor of American students.

Items were accepted as DIF, if item has DIF at level B and C for each of the three methods. Table 8 presents DIF items in booklet 1 according to group, and distributions according to competencies evaluated by PISA 2006 and item formats.

Table 8

Distributions of Items in Booklet 1 that including DIF for Turkish and USA groups according to the item content, measured skill by item and item format

Item #	Items		Competencies	Item Format	Group Favor
4	S213Q01T	Clothes	ISI	CMC	TUR
6	S269Q01	Earth's Temperature	EPS	OR	USA
10	S326Q02	Milk	USE	OR	USA
12	S326Q04T	Milk	EPS	MC	TUR

Table 8 Continue

Item #	Items		Competencies	Item Format	Group Favor
15	S408Q04T	Wild Oat Grass	EPS	CMC	USA
17	S415Q02	Solar Panels	EPS	MC	USA
18	S415Q07T	Solar Panels	ISI	MC	USA
20	S416Q01	The Moon	USE	OR	USA
28	S426Q05	Grand Canyon	EPS	MC	USA
29	S426Q07T	Grand Canyon	ISI	MC	USA
37	S485Q02	Asit Rain	EPS	OR	USA
39	S485Q05	Asit Rain	ISI	OR	USA
41	S493Q03T	Physical Exercise	EPS	MC	TUR
49	S510Q01T	Magnetic Hovertrain	EPS	MC	TUR
56	S527Q01T	Extinction of Dinosaurs	USE	MC	TUR
57	S527Q03T	Extinction of Dinosaurs	EPS	MC	USA

Note. Competencies: ISI = Identify scientific issues, EPS= Explain phenomena scientifically, USE= Use scientific evidence. Item format: OR= Open-constructed response, MC= Multiple-choice, CMC= Complex multiple-choice

Table 8 shows that 16 items in booklet 1 showed DIF, representing27.6% of items in the booklet. Five of the items showing DIF worked in favor of Turkish students while 11 items worked in favor of American students. Table 9 shows DIF items in booklet 5 according to group, and distributions according to competencies and item formats.

Table 9

Distributions of Items in Booklet 5 that including DIF for Turkish and USA groups according to the item content, measured skill by item and item format

Item #	Items		Competencies	Item Format	Group Favor
2	S131Q04T	Good Vibration	ISI	OR	TUR
5	S256Q01	Spoons	EPS	MC	USA
6	S268Q01	Algae	ISI	MC	USA
8	S268Q06	Algae	EPS	MC	USA
12	S304Q03B	Water	EPS	OR	TUR
14	S413Q05	Plastic Age	USE	MC	TUR
16	S416Q01	The Moon	USE	OR	USA
20	S425Q03	Penguin Island	EPS	OR	TUR
23	S428Q01	Bacteria in Milk	USE	MC	TUR
29	S447Q02	Sunscreens	ISI	MC	TUR
33	S458Q01	The Ice Mummy	EPS	MC	TUR
35	S465Q01	Different Climates	USE	OR	USA
36	S465Q02	Different Climates	EPS	MC	USA
39	S466Q05	Forest Fires	USE	MC	TUR

Table 9 Contiue

Item #	Items		Competencies	Item Format	Group Favor
40	S466Q07T	Forest Fires	ISI	MC	USA
41	S477Q02	Mary Montagu	EPS	MC	USA
46	S478Q03T	Antibiotics	EPS	MC	USA
47	S493Q01T	Physical Exercise	EPS	MC	USA
48	S493Q03T	Physical Exercise	EPS	MC	TUR
52	S498Q04	Experimental Digestion	USE	OR	USA
57	S519Q02T	Airbags	EPS	CMC	USA
58	S519Q03	Airbags	ISI	OR	TUR
59	S524Q06T	Penicillin Manufacture	USE	MC	USA
60	S524Q07	Penicillin Manufacture	USE	OR	USA

Note. Competencies: ISI = Identify scientific issues, EPS= Explain phenomena scientifically, USE= Use scientific evidence. Item format: OR= Open-constructed response, MC= Multiple-choice, CMC= Complex multiple-choice

Table 9 shows that 24 items in booklet 5 show DIF, representing40% of items in the booklet. Ten of the items worked in favor of Turkish students while 14 worked in favor of American students. Since two of these items were common in both booklets, it was concluded that 38 items showed DIF.

Possible Source of DIF in Turkish and American Groups. One of the methods used to determine the source of DIF involve sex pert opinion (Ercikan, 2002; Çet, 2006; Bakan Kalaycıoğlu, 2008). A total of 38 items showed DIF, of which 9 were explained at international level. Five science teachers and three assessment experts' opinions were surveyed for these items, results were shown in Table 10.

Table 10

Distribution of Experts' Opinions about the Source of the Bias

Possible Source of Bias	Item content and item number									
	Clothes -1	Grand Canyon -5	Grand Canyon -7	Asit Rain -2	Asit Rain -5	Physical Exercise -1	Physical Exercise -3	Sunscreens -2	Mary Montagu-2	Total number of judgments
Cultural unfamiliarity with the content	xx	xxxxxx x	xxxxxx	xxx		xxxxxx	x		x	26
The word or expression used for the item has different meaning in cultures				x						1

Table 10 Continue

Possible Source of Bias	Item content and item number									
	Clothes -1	Grand Canyon -5	Grand Canyon -7	Asit Rain -2	Asit Rain -5	Physical Exercise -1	Physical Exercise -3	Sunscreens -2	Mary Montagu-2	Total number of judgments
The country groups become more familiar with the item format	xxx			xxx			x	x		8
The skills measured within the item are familiar to the relevant culture		xx	xx		xxxx	x	xxxx		x x x	16
Other							x		x x	3
Total number of judgments	5	9	8	7	4	7	7	1	6	54

Examining Table 10, some of the experts did not express an opinion about all of the items, whereas others provided two possible sources of bias for one item. The most important source of bias was regarded as "cultural unfamiliarity with the content" (26 judgments) and "the skills measured within the item are familiar to the relevant culture" (16 judgments). Another source of bias was regarded as "country groups becoming more familiar with the item format" (8 judgments).

Tables 9 and 10 showed the distributions of items determined as showing DIF according to evaluated competency to determine whether *competency* evaluated in PISA 2006 affected DIF. Examining Tables 9 and 10, it can be seen that 8 of 12 items about using scientific evidence worked in favor of American students; 13 of 20 items about the processes for explaining cases scientifically worked in favor of American students. It was determined that there was no difference between Turkish and American student groups in terms of items for distinguishing scientific situations.

The effect of differences in *item format* on DIF was determined by considering the distributions of item formats showing DIF. According to Tables9 and 10, in both of the booklets, 14 of 24 *multiple choice items* having DIF worked in favor of American students while 10 worked in favor of Turkish students; 9 of 13 *open-ended* items worked in favor of American while 4 worked in favor of Turkish students. Accordingly, although there was not a significant difference between two groups in terms of multiple-choice items, open-ended items provided advantages for American students.

Conclusions and Recommendations

A single factor structure (science literacy) was detected for both booklets following exploratory factor analysis (EFA) of structure equivalence of PISA 2006 science literacy tests conducted in Turkish and American groups. Moreover, the existence of a single factor structure was supported by CFA conducted on data for booklets 1 and 5 completed by Turkish and American groups. Similarly, CFA was used in international studies to determine factor structures between groups and unidimensionality (Gierl, 2000; Ercikan & Kim 2005; Çet, 2006; Yıldırım & Berberoğlu, 2009). Similarly, a previous study of PISA 2003 also presented a single factor structure for Turkish and American groups (Çet, 2006; Yıldırım, 2008; Yıldırım & Berberoğlu, 2009).

MCFA of the test structures showed differences between Turkish and American groups according to culture, factor loads and error variances of items in booklet 1,but factor structures were the same for both groups in terms of inter-factors correlations. Equivalence of "factor values, inter-factors correlation and error variances" of both groups was presented in booklet 5. Consequently, it was decided that the factor structures of both booklets were equivalent in Turkish and American applications of PISA 2006.This finding differs from that of a previous PISA 2003 study (Çet, 2006), which shows difference between translated forms and original form (i.e.,the measured structure was different) between Turkish and American groups.

MH, SIBTEAST and IRT-LR analysis showed that DIF at levels B and C in booklet 1for16 (28%) items, and 24 (40%) items in booklet 5 by all three methods. Of these items, 15 worked in favor of Turkish students while 25 worked in favor of American students. However, since two of these items were common to both booklets, 38 items were found to show DIF in total. Previous studies of Turkish and American data for the PISA 2003 Mathematics literacy test found that different number of items had DIF (Çet, 2006; Yıldırım, 2006; Yıldırım, 2008; Yıldırım & Berberoğlu, 2009).

Expert opinions were sought on 9 items showing DIF in the present study. The expert responses suggested that bias originated in: cultural familiarity, being familiar with the item content and the skills measured by the item. Similarly, cultural difference was reported as a source of bias in large-scale international studies (Gierl & Khaliq; 2001; Ercikan, 2002; Ercikan, Gierl, Mc Creith, Puhan & Koh, 2004).

Among the processes evaluated in the PISA 2006 science literacy test, it was detected that items about differentiating scientific situations and explaining events scientifically were advantageous to the American group compared to the Turkish group, but there was no difference between the groups in items related to usage of scientific evidence. Comparing the two groups according to item formats, two-thirds of the open-ended items showing DIF were found to favor American students. For multiple-choice items, there was a small difference in favor of American students, but this difference was not significant.

The study findings showed that some items in PISA 2006 science literacy tests showed DIF in favor of Turkish students while others favored American students. The results were not of a sufficient scale to affect the average student performance,

but in such international evaluation studies, presenting sources of bias due to descriptive analysis of item scopes will be beneficial for the participant countries where preliminary test of items are conducted.

References

Abbott, M.L. (2007). A Confirmatory Approach to Differential İtem Functioning on an ESL Reading Assessment. Language Testing 2007 24: 7. *Retrieved* January 24 2011 from http://ltj.sagepub.com/content/24/1/7

Acar, T. (2008). *Maddenin Farklı Fonksiyonlaşmasını Belirlemede Kullanılan Genelleştirilmiş Aşamalı Doğrusal Modelleme, Lojistik Regresyon Ve Olabilirlik Oranı Tekniklerinin Karsılaştırılması.* [*Determination of a Differential Item Functioning (DIF) Procedure Using the Hierarchical Generalized Linear Model: A Comparison Study with Logistic Regression and Likelihood Ratio Procedure*] Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Angoff, William,H. (1993). Perspectives on differential item functioning methodology. In Differentional item functioning Eds. (Paul. W. Holland and Howard Wainer). p.3-23.IEA: NJ USA.

Ateşok Deveci, N. (2008). *Üniversiteleraası Kurul Yabancı Dil Sınavının Madde Yanlılığı Bakımından İncelenmesi.* [*Examination of Inter-University Board Foreign Language Test in The Frame of Item Bias]* Yayımlanmamış doktora tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Bakan Kalaycıoğlu, Dilara. (2008). *Öğrenci Seçme Sınavı'nın Madde Yanlılığı Açısından İncelenmesi [Item Bias Analysis of the University Entrance Examination].* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Baker, F. B. (2001). The Basıcs of Item Response Theory. Second edition. ERIC Clearinghouse on Assessment and Evaluation

Bertrand, R. & Boiteau, N. (2003). Comparing the Stability of IRT- Based and non IRT-Based DIF Methods in Different Cultural Contexts Using TIMSS Data. *Educational Resources information center (ERIC)*

Benito, J.G. & Ara, M. J. N. (2000). A Comporison of X^2, RFA ve IRT Based Procedures in the Detection of DIF. *Klıwer Academic Publishers. Netherlands. Quality & iquantity 34:* 17-31

Bolt, S.E. & Ysseldyke, J.E. (2006). Comparingd DIF acrass Math and Reading/Language Arts Tests for Students Receiving a Read – Aloud Accommodation. *Applied Measurement in Education. 19.*(4), 329-355

Camilli, G. & Shepard, A., L., (1994). *Methods for Identifying Biased Test Items.* SAGE Publications. California.

ÇET, S. (2006). *A Multıvarıate Analysıs In Detectıng Dıfferentıally Functıonıng Items Through The Use Of Programme For Internatıonal Student Assessment (Pısa) 2003 Mathematıcs Lıteracy Items.* Yayınlanmamış doktora tezi, ODTÜ, Ankara

Doğan, N. & Öğretmen, T. (2006). Madde Yanlılığını Belirleme Teknikleri Arasında Bir Karşılaştırma. *Eurasian Journal of Educational Research,* 23, pp, 94-105.

Donoghue, J.R, Holland P.W. & Thayer, D.T. (1993). *A Monte Carlo Study of Factors That Affect the Mantel-Haenszel and Standadization Measures of Differential Item Functioning*. Differantial İtem Functioning. Edicted By J.R.Holland and H. Wainer. Lawrence Erlbaum Associaten. London.

Dorans. N.J. & Holland, P.W, (1992). *DIF Detection and Desciription: Mantel-Haenszel and Standardizasion.* In P.W.Holland and H. Wainer (Eds.). Differantial İtem Functioning.Lawrence Erlbaum.

Ellis, B. B., & Raju, N.S.(2003). Test and İtem Bias: What They Are, What They Aren't, and How To Detect Them. *Educational Resources information center (ERIC)*

Embretson, S.E. & Reise, S.T, (2000). *Item Perponse Theory For Psychologists*. London: Lawrance Erlbaum Associates Publishers

Ercikan, K. (2002). Disentangling sources of Differential Item Functioning in Multilanguage Assessments. *International Journal of Testing,* 2(3&4), 199-215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's national achievement tests. Applied Measurement in Education, 17, 301–321. *Retrieved* August 23 2011 from Web: http://dx.doi.org/10.1207/s15324818ame1703_4

Ercikan, K., Mc Creith, T. & Lapointe V. (2005). Factors Associated with Mathematics Achievement and Participation in Advanced Mathematics Courses: An Examination of Gender Differences From an International Perspective. Volume 105(1), *Retrieved* January 8 2007 from Web: http://qix.sagepub.com/cgi/reprint/9/6/859

Ercikan, K. & Kim, K. (2005): Examining the Construct Comparability of the English and French Versions of TIMSS, International Journal of Testing, 5:1, 23-35. *Retrieved* August 19 2011 from Web: http://dx.doi.org/10.1207/s15327574ijt0501_3.

Gierl, M.J.(2000). Construct Equivalance on Translated Achievement Test. *Canadian Journal of Education* 25, 4 (2000), 280-296.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. Journal of Educational Measurement, 38, 164–187.

Gierl, M., Khaliq, S. N. & Boughton, K. (1999). Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications. Paper Presented at the Symposium entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education. Sherbrooke, Québec.

Gotzmann, A.J.(2002). The Effect of Large Ability Differences on Type I Error and Power Rates Using SIBTEST and TESTGRAF DIF Detection Procedures. Paper prepared at the Annual Meeting of American Educational Research Association. New Orleans.

Gotzmann, A., Wright, K. & Rodden, L.(2006). A Comparison of Power Rates for Items Favoring the Reference and Focal group for the Mantel-Haenszel and SIBTEST Procedures. Paper presented at the American Educational Research Association (AERA) in San Francisco, California.

Hambleton, R. K., Clauser, B. E., Mazor,K. M. & Jones, R. W. (1993). Advances in the detection of differentially functioning test items.*European Journal of Psychological Assessment,* 9(1), 1-18.

Hambleton, R. & Rodgers, J. (1995). Item bias review. Practical Assessment, Research & Evaluation, 4(6). Retrieved February 11, 2013 from http://PAREonline.net/getvn.asp?v=4&n=6).

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hambleton, R.K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* California:SAGE Publications.

Jöreskog, K. G. & Sörbon, D. (1993). *LISREL 8: structural equation modeling with the SİMPLİS command language.* Scientific software.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling.* London: Sage.

Le, L.T. (2009). Investigating Gender Differential Item Functioning Across Countries and Test Language of PISA *Science Items.International Journal of Testing,* 9: 122-133.

MEB (2007). *PISA 2006 Ulusal Ön Rapor.* MEB Eğitimi Araştırma ve Geliştirme Dairesi. Ankara.

MEB (2010). *Pısa 2006 Projesi Ulusal Nihai Rapor.* MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı. Ankara.

OECD.(2005). PISA 2003 Data Analysis Manual . PISA 2003 Data Analysis Manual: SPSS® Users

Osterlind, S.J. (1983). *Test Item Bias*. Sage Publications, California

Reynolds, C., Livingston, R.B. & Wilson, V. (2006). *Measurement and Assessment in Education.* Boston: Pearson.

Shepard, L. A., Camili, G. & Williams, D. M. (1985). Validity of Approximation Techniques for Detecting Item Bias. *Journal of Educatıonal Measurement.* Volume 22, No. 2, Summer 1985, pp. 77-105

Skaggs, G. & Lissitz, R.W. (1992) The Consistency of Detecting İtem Bias Across Different Test Administrations: İmplications of Another Failure. *Journal of Educational Measurement Fall* 1992,vol.29, No.3, pp 227-242

Stout, W., Bolt, D. Froelich, A. G., Habing, B, Hartz, S. ve Roussos, L. (2003). Development of a SIBTEST Bundle Methodology for Improving Test Equity With Applications for GRE Test Development. Princeton, NJ: ETS.

Şimşek, Ö.F. (2007). *Yapısal Eşitlik Modellemesine Giriş (Temel İlkeler ve LISREL Uygulamaları)*. Ankara: Ekinoks Yayınevi.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics*. Fifth Edition. Pearson: AB

Thissen, D. (2001). IRTLRDIF v.2.0b: Software For The Computation of The Statistics Involved In Item Response Theory Likelihood-Ratio Tests For Differential Item Functioning.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential Item Functioning Using the Parameters of Item Response Models. In P.W.Holland and H. Wainer (Eds.*). Differantial İtem Functioning*.Lawrence Erlbaum.

Waller, N.G. (2005). EZDIF: A Computer Program For Detecting Uniform And Nonuniform Differential Item Functioning With The Mantel-Haenszel And Logistic Regression Procedures. *Retrieved* September 25 2010 from Web: http://www.psych.umn.edu/faculty/ waller/downloads.htm

Welkenhuysen-Gybels, J. & Billiet, J. (2002). *A Comparison of Techniques for Detecting Cross-Cultural Inequivalence* at the Item Level. Kluwer *Academic* Publishers. 197–218. Netherlands.

Yıldırım, H. H & Berberoğlu, G. (2009). Judgmental and Statistical DIF Analyses of the PISA-2003 Mathematics Literacy Items. International Journal of Testing, 9: 108–121, 2009. *Retrieved* July 17 2011 from http://www.informaworld.com/smpp/title~content=t775653658

Yıldırım, H., Hüseyin. (2006).*The Dıfferentıal Item Functıonıng (Dıf) Analysıs of Mathematıcs Items in The Internatıonal Assessment Programs.* Yayınlanmamış doktora tezi, ODTÜ, Ankara.

Yıldırım, S. (2008). Farklı işleyen maddelerin belirlenmesinde sınırlandırılmış faktör çözümlemesinin olabilirlik–oranı ve Mantel-Haenszel yöntemleriyle karşılaştırılması [Comparison of Restricted-Factor Analysis With Likelihood Ratio and Mantel-Haenszel Methods in DİF Analyses]. *H.Ü.eğitim Fakültesi Dergisi* 34:297-307.

Zhou,j., Gierl, M. J. & Tan, X. (2005).Evaluating the Performance of SIBTEST and MULTISIB Using Different Matching Criteria. *Retrieved* September 25 2010 from *www2.education.ualberta.ca/educ/psych/crame/files/ncme06_jz.pdf*

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

http://pisa2006.acer.edu.au/downloads.php

http://earged.meb.gov.tr

Türk ve Amerikan Örnekleminde PISA 2006 Fen Okuryazarlığı Testindeki Maddelerin Yanlılık Bakımından Araştırılması

Atıf:

Çıkrıkçı Demirtaşlı, N. & Uluştaş, S. (2015). A Study on Detecting of Differential Item Functioning of PISA 2006 Science Literacy Items in Turkish and American Samples. *Eurasian Journal of Educational Research*, 58, 41-60. http://dx.doi.org/10.14689/ejer.2015.58.3

Özet

Problem Durumu: Madde yanlılığı, aynı yetenek düzeyinde oldukları halde bir maddenin doğru yanıtlanma olasılığını, bir gruptaki bireylerin diğer grupta yer alan bireylerden daha az doğru yanıtlama olasılığı bulunmasıdır. Maddenin yanlılık taşıması durumda testle ya da maddeyle, ölçülen özelliğin değeri, sistematik olarak olduğundan daha düşük ya da daha yüksek elde edilir. Bu nedenle test puanlarına dayalı olarak verilecek kararların isabetliliği bakımından test geliştirme ve test uyarlama sürecinde maddelerin olası yanlılık şüphesine karşı sınanması gerekir. Klasik test kuramı (KTK) ve madde tepki kuramına (MTK) göre madde yanlılığı belirlemede farklı yöntemler kullanılmaktadır. Klasik test kuramı çerçevesinde birçok araştırmacı, madde yanlılığını, gruplar arasında madde–aritmetik ortalama ya da madde-test korelâsyonu gibi klasik madde istatistikleriyle karşılaştırma yaparak araştırmaktadır. MTK literatüründe madde yanlılığı kavramı, madde işlev farklılığı (MİF) (DIF:Differential Item Functioning) olarak ifade edilir. Madde yanlılığı analizlerini MTK ile yapmak; bu iki gruptan kestirilen madde parametrelerinin değerlerinin ve bu maddeye ait iki gruptan kestirilen madde karakteristik eğrileri (MKE-Item Characteristic Curve-ICC) arasındaki alanların karşılaştırılmasıdır. Bir test maddesinin madde karakteristik eğrileri referans ve odak gruplar için aynı olmadığında madde her iki grupta aynı biçimde ölçmüyor, diğer bir ifayle MİF gösteriyor demektir. Araştırmalarda genelde bu yöntemlerin birkaçı birlikte kullanılır. Bir testte MİF'in varlığını belirlemek için yapılan araştırmalarda, farklı yöntemlerin kullanıldığı durumlarda yöntemlere göre MİF'li olarak belirlenen maddelerin farklı olduğu görülebilmektedir. Bundan dolayı MİF belirlemek için tek bir yöntem kullanmak yerine birden fazla yöntemi kullanarak araştırma yapmak ve birden fazla yöntemde MİF şüphesi gösteren maddeleri incelemeye almak, yanlı maddelerin belirlenmesinde daha güvenilir sonuç vermektedir. Bu araştırmada da üç farklı yöntem kullanılarak PISA 2006 fen okuryazarlığı testi maddelerinde yanlılık olup olmadığı araştırılmıştır. PISA uygulaması, dünyada politika geliştirenlerin eğitim politikalarını yönlendirmede en çok dikkate aldıkları çalışmalardan biridir. Bu araştırma ile Türkiye'de uygulanan PISA 2006 fen okuryazarlığı testinde yer alan maddelerin herhangi bir yanlılık şüphesi bulundurup bulundurmadığı araştırılmıştır. PISA uygulamalarında kullanılan testlerin orijinali İngilizce dilinde hazırlanmakta ve her katılımcı ülkenin diline çevrilmektedir. Bu nedenle bu tür uygulamalarda maddelerde yanlılığa yol açabilecek en önemli kültürel unsur dildir. Araştırmada PISA 2006 fen okuryazarlığı testini, testlerin hazırlandığı orijinal dil olan İngilizce dilinde alan ülkelerden ABD'nin verileri kullanılmıştır.

Araştırmanın Amacı: Bu araştırmada, PISA 2006 çalışması fen bilimleri okuryazarlığı testi'nin Türk ve ABD öğrenci gruplarında yapı bakımından eşdeğerliğinin incelenmesinin yanı sıra, fen bilimleri okuryazarlığı maddelerinin kültürler arası yanlılık gösterip göstermediği ve varsa olası yanlılık nedenlerinin ortaya konulması amaçlanmıştır.

Araştırmanın Yöntemi: Araştırma 757 Türk ve 856 ABD'li öğrencinin verileri ile gerçekleştirilmiştir. Araştırmada kullanılan fen bilimleri okuryazarlığı testinin Türk ve ABD gruplarında yapı bakımından eşdeğer olup olmadıklarını belirlemek için verilere önce açımlayıcı ve doğrulayıcı faktör analizi uygulanmıştır. Testlerin Türk ve ABD gruplarına göre faktör yapısının kültürlere göre farklılığa sahip olup olmadığını belirlemek için ise çoklu grup doğrulayıcı faktör analizi [(ÇGDFA)(multi group confirmatory factor analysis)] yapılmıştır. PISA 2006 çalışmasında fen bilimleri okuryazarlığı testindeki maddelerde yanlılık olup olmadığının belirlenmesinde ise Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST) ve madde tepki kuramı olabilirlik oranı analizi (MTK-OOA) yöntemleri kullanılmıştır. MİF belirlemede kullanılan yöntemlerin analizleri sonucunda B ve C düzeyinde MİF gösteren maddeler MİF'li olarak alınmıştır.

Araştırmanın Bulguları: Araştırmada ÇGDFA sonuçlarına göre PISA 2006 fen okuryazarlığı testinin Türk ve ABD versiyonlarında her iki kitapçığın faktör yapıları hakkında eşdeğer olduğu kararı verilmiştir. Analizler sonucunda her üç yöntemle ortak olarak B ve C düzeyinde MİF gösteren madde sayısının 1 nolu kitapçıkta 16 (%28), 5 nolu kitapçıkta 24 (%40) olduğu belirlenmiştir. Bu maddelerin 15'i Türk öğrenciler, 25'i ise ABD'li öğrenciler lehine çalışmıştır. Yanlılık kaynağını belirlemek için alınan uzman görüşlerine göre; maddelerde genelde kültüre bağlı olarak, maddenin içeriğine aşina olma ve madde kapsamında ölçülen becerilerin ilgili kültüre tanıdık olma konularının yanlılık kaynağı olduğu ortaya çıkmıştır. PISA 2006 fen okuryazarlığı testinde değerlendirilen süreçlerden, bilimsel durumları ayırt etme ve olguları bilimsel olarak açıklama ile ilgili maddelerin Türk grubuna göre ABD'li gruba avantaj sağladığı belirlenmiştir. Bilimsel kanıtları kullanma ile ilgili maddelerde iki ülke grubu arasında bir farklılık saptanmamıştır. Sonuçta üç yöntemle yapılan analizlerin ortak sonuçlarına göre MİF'li olduğu belirlenen maddelerin madde formatı ve konu alanı açısından hangi gruba avantaj sağladığı çok net olarak ortaya konulmamıştır.

Araştırmanın Sonuçları ve Önerileri: Araştırmanın sonucunda farklı yöntemlerle MİF gösteren maddelerin farklı sayıda olduğu belirlenmiştir. Her üç yöntemle MİF gösterdiği belirlenen 40 madde MİF'li olarak kabul edilmiştir. Uzman görüşlerine göre, bu maddelerden açıklanmış olanlarda, gözlenen olası yanlılık nedenlerinden, kültüre bağlı olarak maddenin içeriğine ve ölçtüğü becerilere aşına olmanın öne çıktığı belirlenmiştir. Maddelerde gözlenen olası yanlılığın nedenlerinin katılımcı ülkelerdeki ortalama öğrenci performansının değerini değiştirecek düzeyde olmadığı sonucuna ulaşılmıştır. Bununla birlikte bu türden uluslararası değerlendirme çalışmalarında maddelerin ön denemelerinin yapıldığı katılımcı ülkelerde, madde kapsamlarının betimsel analizlerle olası yanlılık kaynaklarının ortaya konması yararlı olacaktır.

Anahtar Sözcükler: PISA, madde işlev farklılığı, Mantel-Haenszel, MTK olabilirlik oranı analizi, *SIBTEST*