ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);49-67

Are Differentially Functioning Mathematics Items Reason of Low Achievement of Turkish Students in PISA 2015?*

Serkan ARIKAN**

Abstract
In PISA 2015 the average mathematics score of Turkey decreased dramatically. One of the reasons could be the psychometric properties of mathematics items of PISA 2015. Therefore, it is necessary to evaluate PISA mathematics items for language DIF. In the study, three different DIF detection methods were used: logistic regression (LR), Mantel-Haenszel (MH) and structural equation modeling (SEM). Eleven items were found to have DIF when Turkish and English speaking students were compared. The effect sizes of mathematics performance differences between Turkish and English speaking students before and after excluding DIF items did not change which indicated that DIF items did not cause Turkish students to perform lower than expected. All the DIF items were open response format in which answers were rated by experts and computers. The DIF items favoring Turkish students were mainly related to the basic cognitive process.

Key Words: PISA 2015, Mathematics Performance, DIF, Turkish Student, Low Achievement

INTRODUCTION

The Programme for International Student Assessment (PISA) aims to provide internationally comparable data for 15-year-old students' performance based on reading, mathematics and science. PISA is administered every 3 year which makes possible to monitor progress of educational systems. The results of PISA get great attention by educators, researchers and policy makers as PISA provides detailed information about more than 70 countries. PISA 2015 application had great coverage in which 35 OECD countries and 37 partner countries participated to the assessment. PISA has many additional important features that make it unique and different from other assessments. For instance, PISA links student performance results data with student level variables like students' background and attitudes towards learning and with school level variables like school characteristics. PISA aims to measure students' capacity to apply knowledge and skills in key subjects which is defined as "literacy". (OECD, 2016a).

Turkey, a member of OECD, participates PISA regularly since 2003. Turkey's performances on mathematics were below the average score of 500; 423 in PISA 2003, 424 in PISA 2006, 445 in PISA 2009, 448 in PISA 2012, and 420 in PISA 2015. Similarly the average science scores of Turkey were 434 in PISA 2003, 424 in PISA 2006, 454 in PISA 2009, 463 in PISA 2012, and 425 in PISA 2015; the average reading scores of Turkey were 441 in PISA 2003, 447 in PISA 2006, 464 in PISA 2009, 475 in PISA 2012, and 428 in PISA 2015 (MEB, 2015; MEB, 2016). Through PISA 2012, Turkey had an increasing trend in the scores, however, in PISA 2015 the average scores decreased dramatically. The reasons of this very low score on PISA 2015 are necessary to be investigated.

There might be several reasons of the low scores of Turkish students in PISA 2015. There might be a problem in psychometric properties of items that were used in the PISA 2015 assessment; there might be a problem in the comparability of the samples over years; the change in test administration method (computer based administration instead of paper and pencil test) might cause lower scores. It

*An early draft of this paper was presented at European Conference on Educational Research (ECER) in Bolzano, Italy in 2018.

**Asst. Prof. Dr., Mugla Sitki Kocman University, Faculty of Education, Mugla-Turkey, serkanarikan@mu.edu.tr, ORCID ID: 0000-0001-9610-5496

To cite this article:
Arıkan, S. (2019). Are differentially functioning mathematics items reason of low achievement of Turkish students in PISA 2015?. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 49-67. DOI: 10.21031/epod.466860

Received: 03.10.2018
Accepted:12.02.2019

is also possible that the low scores might be as a result of the change in the curriculum, educational practices or country level educational policies in Turkey. This study focused on the psychometric properties of the PISA 2015 mathematics items as a source of low scores of Turkish students.

Comparative assessments should be fair to all groups of students. Psychometric properties of these assessments should be controlled to prevent any unintended bias. PISA is mainly developed in English first and then adapted to other languages including Turkish (OECD, 2017). Therefore, it is necessary to evaluate whether PISA mathematics items functioned differently for Turkish and English speaking students who answered adapted items and original items, respectively. Finding evidence for fairness of items in terms of psychometric properties could help us to eliminate one of the possible reasons of sharp decrease of Turkish students' mathematics performance in 2015.

Differential item functioning (DIF) detection methods are widely used to evaluate the fairness and equality of tests on item level in investigating the comparability of translated and/or adapted measures (Zumbo, 2007). DIF occurs and threatens the comparability of scores if students with the similar ability level on the underlying construct in different groups do not have the similar probability of getting the right answers for a specific item (van de Vijver & Leung, 1997; Zumbo, 2007). Evaluating items in terms of DIF is a necessary preliminary analysis before conducting any comparative study. Otherwise, if a test contains items having DIF, observed differences in scores could be related to the problematic items rather than true differences on the underlying trait or ability (He & van de Vijver, 2013). If an item is detected as having DIF statistically, the context of the item should be examined by experts to evaluate whether the item indeed biased against one group systematically (van de Vijver & Leung, 1997). However, judgmental expert evaluation alone might not be always successful to detect why DIF occurs. For example, Angoff (1993) reported that even item writers often had problems to understand why some perfectly reasonable items showed large amounts of DIF. Some scholars investigated whether student background variables could be potential explanations of sources of DIF (Joldersma & Bowen, 2010; Liu et al., 2016; Zumbo & Gelin, 2007).

PISA items are prepared very carefully under the guidance of the experts by international team of item developers. Translatability reviews are conducted considering translation, adaptation and cultural issues (OECD, 2017). However, many researchers reported that PISA mathematics items contained DIF (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009). Yildirim and Berberoglu (2009) reported that 5 out of 21 mathematics items in PISA 2003 were flagged as having DIF in the comparison of Turkish and American students (3 of these items favored Turkish students). Lyons-Thomas et al. (2014) found that there were gender DIF in PISA 2009 mathematics items of students in Canada, Finland, Shanghai, and Turkey. Demir and Kose (2014) identified many DIF items in PISA 2009 mathematics assessment when they compare answers of Turkish students with German, Finish and American students. Therefore, there is a possibility that PISA 2015 mathematics items might contain DIF items that could cause a decline in Turkish students' mathematics scores. There is not any study that investigated whether PISA 2015 items contained DIF across Turkish and English speaking students.

Purpose of the Study

Having DIF items for a language group is a threat to comparability of test scores. In this study, PISA 2015 mathematics items were analyzed in terms of DIF for Turkish, English and American students. The main idea is that whether the low mathematics scores of Turkish students could be due to DIF items against Turkish students. Therefore, in order to test this claim, DIF analyses using answers of Turkish and English student, as well as Turkish and American students were conducted separately. The research questions of this study were

(1) Are there any items having DIF in PISA 2015 mathematics test in comparing Turkish and English students?

(2) Are there any items having DIF in PISA 2015 mathematics test in comparing Turkish and American students?

(3) Are there any changes in the effect sizes of mathematics performance differences among groups before and after excluding DIF items, if any?

METHOD

Participants

The data of this study were obtained from the PISA 2015 data set. In PISA, the target population is all 15-year-old students of participating countries. PISA has rotated booklet design in which each student answers linked portion of all items. Therefore, ability level of each student could be estimated from all items without requiring a student to answer all items (OECD, 2016b). This study used the data of all Turkish, English and American students who answered mathematics items in booklets 43, 45, and 47. These three booklets were selected because they included all the items and there were no overlap of items. The participants were 491 Turkish students, 1154 English students and 448 American students.

Instrument

In PISA 2015, a total of 69 mathematics items were used to collect information about students' mathematics performance and a student responded approximately 23 mathematics items. PISA aims to measure mathematical literacy level of students defined as the capacity of students to apply acquired knowledge and skills to different problems and challenges they encounter. The mathematical processes measured in PISA are formulate (formulating situations mathematically), employ (employing mathematical concepts, facts, procedures and reasoning), and interpret (interpreting, applying and evaluating mathematical outcomes) (OECD, 2016b). These mathematical processes have a hierarchical order in which interpret represents the highest cognitive process. In Table 1, Table 2 and Table 3, item number, item code, item label, item format, cognitive processes measured by each item and item-level percentage correct values for Turkish, English and American students in booklet 43, 45 and 47 were reported.

Data Analysis

In the study, three different DIF detection methods were used. These DIF detection methods were logistic regression (LR), Mantel-Haenszel (MH) and structural equation modeling (SEM). As each DIF method is based on different statistical procedures, and studies reported that there might be low to medium coherence among DIF detection methods (Atalay, Gok, Kelecioglu & Arsan, 2012), more than one method was used. In order to get more consistent findings, an item that showed DIF in at least two different methods was considered to contain DIF across language groups. Sixty-nine mathematics items were evaluated in terms of DIF for Turkish-English and Turkish-American students groups.

In the logistic regression method, as the first step, only total score (model 1), then total score and grouping variable (model 2), and finally total score, grouping variable and their interaction (model 3) were used as predictors. Significance of country and their interaction, and the change in R^2 value were taken as evidence for uniform bias and non-uniform bias, respectively (Zumbo, 1999). Zumbo and Thomas (1997) proposed that ΔR^2 (the difference between model 3 and model 1) higher than 0.130 indicates moderate DIF and higher than 0.260 indicates large DIF. Jodoin and Gierl (2001) proposed lower values to detect DIF; ΔR^2 higher than 0.035 indicates moderate DIF and higher than 0.070 indicates large DIF. In this study the criteria of Jodoin and Gierl was used to detect DIF items as it requires lower values which allows to detect more items. Therefore, the possibility to omit an

item that might have bias will be minimized. SPSS 22.0 programs were used to conduct logistic regression analysis.

Table 1. Item Descriptions for Booklet 43

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B43_1	CM033Q01S	A View Room-Q01	SMC	interpret	.56	.74	.75
B43_2	CM474Q01S	Running Time-Q01	SMC	employ	.44	.63	.64
B43_3	DM155Q02C	Population Pyramids-Q02	OR	interpret	.22	.57	.43
B43_4	CM155Q01S	Population Pyramids-Q01	CMC	employ	.46	.66	.63
B43_5	DM155Q03C	Population Pyramids-Q03	OR	employ	.07	.13	.14
B43_6	CM155Q04S	Population Pyramids-Q04	CMC	interpret	.32	.54	.43
B43_7	CM411Q01S	Diving-Q01	OR	employ	.25	.52	.43
B43_8	CM411Q02S	Diving-Q02	SMC	interpret	.29	.48	.51
B43_9	CM803Q01S	Labels-Q01	OR	formulate	.10	.28	.20
B43_10	CM442Q02S	Braille-Q02	CMC	interpret	.14	.20	.25
B43_11	DM462Q01C	Third Side-Q01	OR	employ	**.13**	**.01**	**.03**
B43_12	CM034Q01S	Bricks-Q01	OR	formulate	.17	.32	.23
B43_13	CM305Q01S	Map-Q01	SMC	employ	.31	.39	.42
B43_14	CM496Q01S	Cash Withdrawal-Q01	CMC	formulate	.23	.47	.41
B43_15	CM496Q02S	Cash Withdrawal-Q02	OR	employ	.47	.68	.59
B43_16	CM423Q01S	Tossing Coins-Q01	SMC	interpret	**.77**	.84	**.71**
B43_17	DM406Q01C	Running Tracks-Q01	OR	employ	**.09**	.24	**.07**
B43_18	DM406Q02C	Running Tracks-Q02	OR	formulate	.01	.08	.04
B43_19	CM603Q01S	Number Check-Q01	CMC	employ	.23	.32	.31
B43_20	CM571Q01S	Stop The Car-Q01	SMC	interpret	.22	.39	.34
B43_21	CM564Q01S	Chair Lift-Q01	SMC	formulate	**.39**	**.37**	.41
B43_22	CM564Q02S	Chair Lift-Q02	SMC	formulate	.33	.42	.35

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Table 2. Item Descriptions for Booklet 45

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B45_1	CM447Q01S	Tile Arrangement-Q01	SMC	employ	.52	.55	.53
B45_2	CM273Q01S	Pipelines-Q01	CMC	employ	**.37**	.37	**.32**
B45_3	CM408Q01S	Lotteries-Q01	CMC	interpret	.29	.40	.34
B45_4	CM420Q01S	Transport-Q01	CMC	interpret	.30	.54	.51
B45_5	CM446Q01S	Thermometer Cricket-Q01	OR	formulate	.65	.68	.67
B45_6	DM446Q02C	Thermometer Cricket-Q02	OR	formulate	.02	.08	.05
B45_7	CM559Q01S	Telephone Rates-Q01	SMC	interpret	**.54**	.59	**.49**
B45_8	DM828Q02C	Carbon Dioxide-Q02	OR	employ	.52	.66	.57
B45_9	CM828Q03S	Carbon Dioxide-Q03	OR	employ	.24	.29	.27
B45_10	CM464Q01S	Fence-Q01	OR	formulate	**.20**	**.19**	**.15**
B45_11	CM800Q01S	Computer Game-Q01	SMC	employ	**.88**	**.86**	**.78**
B45_12	CM982Q01S	Employment Data-Q01	OR	employ	.71	.84	.81
B45_13	CM982Q02S	Employment Data-Q02	OR	employ	.14	.40	.35
B45_14	CM982Q03S	Employment Data-Q03	CMC	interpret	.57	.63	.64
B45_15	CM982Q04S	Employment Data-Q04	SMC	formulate	.31	.49	.37
B45_16	CM992Q01S	Spacers-Q01	OR	formulate	.48	.70	.68
B45_17	CM992Q02S	Spacers-Q02	OR	formulate	.06	.11	.10
B45_18	DM992Q03C	Spacers-Q03	OR	formulate	**.05**	**.03**	.05
B45_19	CM915Q01S	Carbon Tax-Q01	SMC	employ	.31	.49	.39
B45_20	CM915Q02S	Carbon Tax-Q02	OR	employ	.54	.66	.61
B45_21	CM906Q01S	Crazy Ants-Q01	SMC	employ	.35	.61	.47
B45_22	DM906Q02C	Crazy Ants-Q02	OR	employ	.18	.39	.31
B45_23	DM00KQ02C	Wheelchair Basketball-Q02	OR	formulate	.02	.09	.05

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Table 3. Item Descriptions for Booklet 47

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B47_1	CM909Q01S	Speeding Fines-Q01	OR	interpret	.48	.90	.84
B47_2	CM909Q02S	Speeding Fines-Q02	SMC	employ	.20	.46	.51
B47_3	CM909Q03S	Speeding Fines-Q03	OR	interpret	.06	.24	.26
B47_4	CM949Q01S	Roof Truss Design-Q01	CMC	employ	.38	.67	.60
B47_5	CM949Q02S	Roof Truss Design-Q02	CMC	employ	.20	.33	.26
B47_6	DM949Q03C	Roof Truss Design-Q03	OR	formulate	.18	.24	.30
B47_7	CM00GQ01S	Advertising Column-Q01	OR	formulate	**.05**	.06	**.03**
B47_8	DM955Q01C	Migration-Q01	OR	interpret	.41	.79	.68
B47_9	DM955Q02C	Migration-Q02	OR	interpret	**.34**	**.30**	**.21**
B47_10	CM955Q03S	Migration-Q03	OR	employ	.01	.08	.05
B47_11	DM998Q02C	Bike Rental-Q02	OR	interpret	.52	.77	.84
B47_12	CM998Q04S	Bike Rental-Q04	CMC	employ	.28	.30	.28
B47_13	CM905Q01S	Tennis balls-Q01	CMC	interpret	.50	.70	.72
B47_14	DM905Q02C	Tennis balls-Q02	OR	interpret	.20	.41	.31
B47_15	CM919Q01S	Fan Merchandise-Q01	OR	employ	.69	.83	.75
B47_16	CM919Q02S	Fan Merchandise-Q02	OR	formulate	.21	.39	.40
B47_17	CM954Q01S	Medicine doses-Q01	OR	employ	.36	.64	.70
B47_18	DM954Q02C	Medicine doses-Q02	OR	employ	.13	.35	.33
B47_19	CM954Q04S	Medicine doses-Q04	OR	employ	.01	.29	.21
B47_20	CM943Q01S	Arches-Q01	SMC	formulate	.37	.45	.47
B47_21	CM943Q02S	Arches-Q02	OR	formulate	.00	.02	.01
B47_22	DM953Q02C	Flu test-Q02	OR	interpret	.11	.33	.31
B47_23	CM953Q03S	Flu test-Q03	OR	formulate	.12	.47	.38
B47_24	DM953Q04C	Flu test-Q04	OR	formulate	.00	.11	.07

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

The Mantel-Haenszel DIF detection method is based on building of *K* two-by-two contingency tables, where K represents the number of discrete score categories that are used to match the comparison groups. For each matched score level, the expected and observed ratios are compared by chi-square method (Holland & Thayer, 1986). Then The MH D-DIF index is calculated using these comparisons with logarithmic transformations in which a negative value indicates the item favors reference group over the focal group (Holland & Thayer, 1988). Educational Testing Service (ETS) proposed a criterion to flag DIF items: The MH D-DIF index between 1 and 1.5 indicates moderate DIF and The MH D-DIF index higher than 1.5 indicated large DIF (Zieky, 1993). DIFAS 5.0 program was used for MH DIF detection analysis (Penfield, 2005).

In the SEM procedure, a Confirmatory Factor Analysis (unifactorial, with all items as indicators of the latent variable) is conducted to assess configural, metric and scalar invariance. The difference between incremental types of model fit is evaluated as the factor loadings and intercepts are forced to be equal for comparison groups (van de Vijver, 2017). If the difference in comparative fit index (CFI) and Tucker Lewis index (TLI) between configural, metric and the scalar invariance model is larger than 0.010, the modification indices are investigated to identify DIF items (Cheung and Rensvold, 2002). Mplus 7.4 program was used for SEM DIF detection procedure (Muthen & Muthen, 2015).

After detecting DIF items, the effect sizes of mathematics performance differences among student groups before and after excluding DIF items were calculated. The change in effect sizes was investigated. Effect size allows researchers to compare the difference between groups without being affected from sample size (Field, 2013). For comparing means of two groups, Cohen's *d* is frequently used as an indicator of effect size. Cohen's *d* is calculated as the difference between the group means divided by the pooled standard deviation. Cohens' *d* value around 0.2 is considered as a

small, around 0.5 represents a moderate and around 0.8 is considered as large effect size (Cohen, 1988).

RESULTS

Preliminary Analysis

Reliability Analysis of the Instrument

Cronbach's alpha reliability coefficients of the PISA 2015 mathematics tests for booklets 43, 45 and 47 were calculated as 0.78, 0.79, 0.76 for Turkish students, 0.81, 0.84, 0.85 for English students, and 0.80, 0.86, 0.86 for American students, respectively. These values indicated good internal consistency (Cicchetti, 1994).

DIF Results

In this section, results based on LR, MH and SEM DIF detection methods were presented. Overall results were compared at the end of this section.

Logistic Regression DIF Results

DIF results using LR method was presented in Table 4. In comparing answers of Turkish and English student, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_8, B47_9 and B47_19) were flagged as having DIF. When answers of Turkish and American student were compared, 14 out of 69 items (B43_11, B43_15, B43_16, B45_10, B45_11, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF.

Table 4. Logistic Regression DIF Results

Item No	Booklet 43		Booklet 45		Booklet 47	
	TR-ENG ΔR2	TR-USA ΔR2	TR-ENG ΔR2	TR-USA ΔR2	TR-ENG ΔR2	TR-USA ΔR2
1	.012	.027	.014	.016	**.089****	**.064***
2	.024	.012	.014	.020	.001	.015
3	.033	.012	.008	.006	.000	.009
4	.008	.019	.014	.028	.003	.000
5	.002	.006	.011	.014	.013	.017
6	.007	.000	.003	.006	**.046***	**.039***
7	.004	.004	.003	.018	**.047***	**.057***
8	.010	.029	.005	.008	**.041***	.031
9	.004	.003	.016	.014	**.107****	**.194****
10	.017	.001	**.052***	**.059***	.009	.016
11	**.299****	**.147****	.030	**.094****	.005	**.043***
12	.001	.006	.005	.014	.006	.010
13	.005	.002	**.038***	**.053***	.000	.004
14	.003	.018	.005	.002	.014	**.048***
15	.003	**.036***	.002	.002	.011	.033
16	.011	**.045***	.013	.033	.009	.004
17	.012	.031	.009	.006	.000	.025
18	.011	.029	**.118****	**.121****	.001	.004
19	.003	.005	.003	.000	**.039***	**.056***
20	.010	.015	.018	.012	.005	.003
21	.021	.009	.015	.006	.001	.001
22	.008	.012	.013	.013	.005	.000
23	-	-	.013	.009	.019	.014
24	-	-	-	-	.014	.022

Note: * indicates the item shows moderate level of DIF; ** indicates the item shows large level of D

Mantel-Haenszel DIF Results

DIF results using MH method was presented in Table 5. In comparing answers of Turkish and English student, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_10 and B47_19) were flagged as having DIF. When answers of Turkish and American student were compared, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF.

Table 5. Mantel-Haenszel DIF Results

Item No	Booklet 43		Booklet 45		Booklet 47	
	TR-ENG ΔMH	TR-USA ΔMH	TR-ENG ΔMH	TR-USA ΔMH	TR-ENG ΔMH	TR-USA ΔMH
1	-.215	-.566	.444	.306	**-1.634****	**-1.445***
2	-.212	-.390	.579	.630	-.166	-.648
3	-.969	-.504	.068	.312	-.2486	-.519
4	.168	-.072	-.551	-.951	-.0888	-.010
5	.124	-.263	.605	.198	.634	.539
6	-.306	-.041	-.642	-.573	**1.495***	.441
7	-.386	-.151	.142	.481	**1.196***	**1.850****
8	-.186	-.516	-.130	.074	-.945	-.403
9	-.561	-.124	.594	.273	**2.260****	**2.241****
10	.947	.102	**1.843****	**1.611****	**-1.057***	NA
11	**3.910****	**2.732****	.812	**1.107***	-.297	**-1.106***
12	-.030	.341	-.355	-.611	.095	.162
13	.109	-.213	**-1.078***	**-1.131***	-.036	-.212
14	-.262	-.275	.310	-.049	.755	**1.564****
15	.149	.259	-.181	.162	.468	.980
16	.168	1.040	-.593	-.893	.433	.219
17	-.306	.878	.591	.259	.108	-.617
18	-.802	-.616	**3.385****	NA	-.095	.207
19	.204	.018	-.215	-.160	**-3.060****	**-1.820****
20	-.132	-.184	-.024	.006	.318	.099
21	.678	.306	-.681	-.201	NA	NA
22	.112	.339	-.540	-.450	-.263	-.068
23	-	**-**	-.751	-.071	-.923	-.421
24	-	**-**	**-**	**-**	NA	NA

Note: * indicates the item shows moderate level of DIF; ** indicates the item shows high level of DIF; NA indicates calculation problem due to low correct response ratio

SEM DIF Results

SEM DIF results were presented in Table 6. In comparing answers of Turkish and English student, 4 out of 69 items (B45_2, B45_10, B45_13 and B45_18) were flagged as having DIF. When answers of Turkish and American student were compared, 2 out of 69 items (B45_13 and B47_9) were flagged as having DIF.

Table 6. SEM DIF Results

Booklet	Model	χ^2/df	RMSEA	CFI	ΔCFI	TLI	ΔTLI	DIF ITEMS
43 TR-UK	Configural	1.192**	.027	.971		.967		None
	Metric	1.222**	.029	.966	.005	.962	.005	
	Scalar	1.232**	.029	.963	.003	.961	-.001	
43 TR-USA	Configural	1.140*	.030	.962		.958		None
	Metric	1.159*	.032	.957	.005	.952	.006	
	Scalar	1.162*	.032	.954	.003	.951	.001	
45 TR-UK	Configural	1.221**	.028	.967		.963		
	Metric	1.220**	.028	.967	.000	.964	-.001	
	Scalar	1.342***	.035	.946	.021	.943	.021	2, 10, 13, 18
	Scalar-Items Removed	1.309***	.033	.957	.010	.954	.010	
45 TR-USA	Configural	1.159*	.032	.960		.957		13
	Metric	1.158*	.032	.961	-.001	.957	.000	
	Scalar	1.199**	.036	.948	.013	.945	.012	
	Scalar-Items Removed	1.180**	.034	.957	.004	.954	.003	
47 TR-UK	Configural	1.558***	.045	.940		.934		None
	Metric	1.539***	.044	.939	.001	.936	-.002	
	Scalar	1.635***	.048	.929	.010	.925	.011	
	Scalar-Items Removed	1.621***	.048	.930	.009	.926	.010	
47 TR-USA	Configural	1.511***	.057	.901		.891		9
	Metric	1.549***	.059	.889	.013	.883	.008	
	Scalar	1.577***	.061	.883	.006	.877	.006	
	Scalar-Items Removed	1.531***	.058	.893	-.004	.887	-.004	

*$p < .05$. **$p < .01$. ***$p < .001$.

Overview of DIF Results

Since each DIF detection method is based on different calculations, an item flagged by at least two method was considered as containing DIF (Table 7). In comparing answers of Turkish and English student, 9 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9 and B47_19) were flagged as having DIF by at least two methods. It is necessary to report which items favored Turkish students and which items favored English students. Among these 9 items, 6 of them favored Turkish students (B43_11, B45_10, B45_18, B47_6, B47_7, B47_9) whereas 3 of them favored English students (B45_13, B47_1, B47_19).

When answers of Turkish and American student were compared, 10 out of 69 items (B43_11, B45_10, B45_11, B45_13, B47_1, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF by at least two methods. Among these 10 items, 5 of them favored Turkish students (B43_11, B45_10, B47_7, B47_9, B47_14) whereas 4 of them favored American students (B45_13, B47_1, B47_11, B47_19). LR results suggested that item B45_11 had non-uniform DIF. The related graphical percentages were given in Appendix A and B. The flagged items were generally consistent across Turkish-English and Turkish-American student comparisons. Items B43_11, B45_10, B47_7 and B47_9 favored Turkish students whereas items B45_13, B47_1, B47_19 favored English speaking students.

Table 7. Overall DIF Results

Booklet	LR	MH	SEM	Items Commonly Flagged
43 TR-UK	11	11	-	11^{TR}
43 TR-USA	11, 15, 16	11	-	11^{TR}
45 TR-UK	10, 13, 18	10, 13, 18	2, 10, 13, 18	10^{TR}, 13^{UK}, 18^{TR}
45 TR-USA	10, 11, 13, 18	10, 11, 13	13	10^{TR}, 11^{*}, 13^{USA}
47 TR-UK	1, 6, 7, 8, 9, 19	1, 6, 7, 9, 10, 19	-	1^{UK}, 6^{TR}, 7^{TR}, 9^{TR}, 19^{UK}
47 TR-USA	1, 6, 7, 9, 11, 14, 19	1, 7, 9, 11, 14, 19	9	1^{USA}, 7^{TR}, 9^{TR}, 11^{USA}, 14^{TR}, 19^{USA}

Note: TR: items favoring Turkish students; UK: items favoring English students; USA: items favoring American students; * non-uniform DIF

Table 8 showed item formats and cognitive domains measured by the DIF items. All the DIF items were open response format in which students constructed the answers and then the answers were rated. Also, among 7 items that favored Turkish students 4 of them were related to formulate cognitive process which is the lowest cognitive process in PISA mathematics assessment. There was no formulate items that favored English or American students.

Table 8. Item Characteristics of DIF Items

Item No	Favoring	Item Label	Item Format	Cognitive Domain
B43_11	Turkish	Third Side - Q01	OR	Employ
B45_10	Turkish	Fence - Q01	OR	Formulate
B45_18	Turkish	Spacers - Q03	OR	Formulate
B47_6	Turkish	Roof Truss Design - Q03	OR	Formulate
B47_7	Turkish	Advertising Column - Q01	OR	Formulate
B47_9	Turkish	Migration - Q02	OR	Interpret
B47_14	Turkish	Tennis balls - Q02	OR	Interpret
B45_13	English&American	Employment Data - Q02	OR	employ
B47_1	English&American	Speeding Fines - Q01	OR	interpret
B47_11	American	Bike Rental - Q02	OR	interpret
B47_19	English&American	Medicine doses - Q04	OR	employ

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Effects of DIF Items on Mathematics Performance Differences

There were mathematics performance differences between Turkish students and English speaking students. Effect size, the standardized mean-difference, allows us to compare the difference between groups without being affected from sample size (Field, 2013). In this part, the original effect sizes and the effect sizes excluding DIF items were reported (Table 9). Between Turkish and English students, there were .51 to .93 effect size differences originally in these booklets. According to Cohen (1988), these values represent moderate to large difference between students. When all DIF items were excluded, effect sizes did not change. Similarly, between Turkish and American students, the original effect sizes were calculated as .28 to .85. According to Cohen (1988), these values represent small to large difference between students. When all DIF items were excluded, effect sizes were very close. The evaluation of the effect size change implied that DIF items generally balanced out each other and did not create any disadvantageous results for Turkish students.

Table 9. Effect Size Change

Booklet	43 TR-UK	43 TR-USA	45 TR-UK	45 TR-USA	47 TR-UK	47 TR-USA
Effect Size	.74	.53	.51	.28	.93	.85
All Items						
Effect Size	.78	.57	.51	.29	.94	.84
Excluding all DIF Items						
Effect Size	.78 Item11	.57 Item11	.54 Item10	.30 Item10	.86 Item1	.80 Item1
Excluding a DIF Item						
Effect Size			.48 Item13	.30 Item11	.96 Item6	.86 Item7
Excluding a DIF Item						
Effect Size			.53 Item18	.24 Item13	.94 Item7	.93 Item9
Excluding a DIF Item						
Effect Size					.99 Item9	.80 Item11
Excluding a DIF Item						
Effect Size					.91 Item19	.88 Item14
Excluding a DIF Item						
Effect Size						.83 Item19
Excluding a DIF Item						

Note: Item numbers given in the table represents the eliminated items.

DISCUSSION

This study has a great importance as it aimed to shed a light on possible causes of low mathematics scores of Turkish students in PISA 2015. Through PISA 2012, Turkey had an increasing trend in their mathematics scores, however, in PISA 2015 the average mathematics score decreased dramatically. In the study, whether the low performance of Turkish students could be due to differentially functioning items was investigated. As PISA is mainly developed in English first and then adapted to other languages including Turkish, evaluating whether PISA mathematics items functioned differently for Turkish and English speaking students was the main focus of the study. In comparing responses of Turkish and English students, 9 items (out of 69) were detected as having DIF. Similarly, 10 items were found to have DIF when Turkish and American students were compared. The surprising finding was that among these DIF items, more items favored Turkish students than they favored English or American students. The standardized mathematics performance differences (measured by effect-size) between Turkish and English speaking students before and after excluding DIF items did not change. Therefore, it is concluded that DIF items did not cause Turkish students to perform lower. Therefore, there is no evidence that PISA items created a disadvantage for Turkish students. Therefore, among possible reasons of low achievement of Turkish students, a problem due to the psychometric properties of PISA items was eliminated. There is still a further need to investigate and focus on other possible reasons of low achievement of 15-year-old Turkish students by conducting new comparative studies.

The possible reasons of these lower scores in PISA 2015 could be the problem of comparability of the Turkish samples over years; the effects of change in test administration method (computer based administration instead of paper and pencil test); the change in the curriculum, educational practices or country level educational policies. One of the reasons of the decrease in the PISA scores could be the selected sample of Turkey. The sampling procedure and coverage rates were reported in PISA technical reports. The coverage rates are important as they give clues about the representativeness of the population. Turkey's coverage rates in PISA were increased over years. The coverage rates were 36% in 2003, 47% in 2006, 57% in 2009, 68% in 2012 and 70% in 2015. Spaull (2017) studied coverage rates and sample of Turkey and he concluded that there was a large change in the proportions of Turkish students that were not sampled in PISA, therefore the validity of the comparisons of the results could have some problems. There is a need to conduct further studies on these sampling issue of Turkey. The other reason could be the change in the administration method of PISA. There was a shift from paper-and-pencil tests in PISA 2012 to computer-based tests (CBT) in 2015. There is a debate over effect of CBT on test results (Jerrim, 2016; Jerrim, Micklewright,

Heine, Salzer, & McKeown, 2018; Komatsu & Rappleye, 2017). Investigating possible effects of CBT on Turkish students' scores would be an informative study about the decrease in scores. Another reason of the decrease in the scores could be related to curriculum change and educational policies. Students who took PISA 2015 in Turkey were mainly 9th or 10th graders. In Turkey, there are frequent changes in curriculum and educational policies in all level of educational system. For instance, in 2012, when students who join the PISA 2015 administration were in 6th or 7th grade, the K-12 education system in Turkey has undergone some major changes and students were allowed to continue their high school in the form of distant education (Gün & Baskan, 2014). The effects of these curriculum and system changes on PISA scores are worth to investigate. The last but not the least, the congruence between educational practices in Turkey and cognitive skills measured in PISA might create a low score for Turkish students. As PISA aims to measure students' capacity to apply knowledge and skills that are related to be successful in modern societies (OECD, 2016a), acquiring curriculum related knowledge might not be enough to be successful in PISA. However, in TIMSS 2015, another large scale assessment that focus more on curriculum, Turkish students increased their scores in both mathematics and science (Yıldırım et al., 2016). A study focuses on the increase of scores on the curriculum focused large scale assessment but the decrease of scores on capacity focused large scale assessment of Turkish students would be informative.

This study found DIF items in mathematics assessment, however the DIF items did not lead Turkish students to perform lower in PISA 2015. The DIF flagged items were generally consistent across Turkish-English and Turkish-American student comparisons. Among 9 items that were flagged as DIF in Turkish and English student comparison, 7 of them were also flagged in Turkish-American comparison. As these items were not released, it was not possible to evaluate the content of items to speculate why these items contained DIF consistently across different comparison groups. There is a need to identify possible sources of DIF, hopefully after items are released. The results of the study were consistent with the other researchers who found DIF items in PISA (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009).

Although mathematics items were not released, there was an information about item format and cognitive processes measured by each item. There were relationship between DIF items and their format and cognitive processes. First of all, all the DIF items were open response items in which students' answers were rated by experts or computers (OECD, 2017). Among 69 items, 18 open response items were coded by experts and 22 open response items were coded by computers. Multiple coding design was used to monitor coder reliabilities within and across countries. The open-ended coding system was used to simplify the coding process. National Project Managers of each country were expected to investigate the systematic pattern of irregularities. For OECD countries, the median within-country agreement of raters was 97.5% and the median across-country agreement of raters was 97.9% in mathematics. For Turkey, within-country agreement of raters was 97.7% and across-country agreement of raters was 93.9% which was the second lowest (OECD, 2017). As all DIF items were open response items, and across-country agreement of Turkey was lower than OECD countries, it would be informative to know whether the coding could cause an advantage or disadvantage for Turkish students. Another issue is that the DIF items favoring Turkish students were mainly related to formulate cognitive process. Formulate cognitive process is defined as formulating situations mathematically which is the lowest cognitive process in PISA. In Turkish educational system there are problems that teachers do not give adequate emphasis to develop higher cognitive processes. Turkish students generally encounter with items that are related to basic skills as comprehension rather than higher order thinking skills as problem solving (Arikan, van de Vijver & Yagmur, 2016; Doganay & Bal, 2010; Temur, 2012). Therefore, Turkish students' high familiarity of basic cognitive skills could cause more formulate items to be detected as having DIF.

In the study three different DIF identification methods were applied. Logistic regression and Mantel-Haenszel DIF methods gave similar results compared to structural equation modeling DIF method. Structural equation modeling DIF results were more conservative in detecting items as DIF compared to the two other methods. Although logistic regression and Mantel-Haenszel methods

produced similar results, logistic regression method detected more items as having DIF compared to Mantel-Haenszel method. Except one item in booklet 47 (TR-UK comparison), all items flagged by Mantel-Haenszel were also flagged by logistic regression method in all booklets for all comparisons. Therefore, in this study, it was observed that logistic regression method flagged more items as having DIF. On the other hand, structural equation modeling DIF method flagged items having DIF very rarely compared to other two methods. Atalay et al. (2012) compared logistic regression and Mantel-Haenszel methods in their simulation study and concluded that Mantel-Haenszel method was more sensitive in detecting DIF items. On contrary to this study, Gok, Kelecioglu and Dogan (2010) found more gender and school type DIF using Mantel-Haenszel method compared to logistic regression method in high school entrance examination items of Turkey. These findings indicate that different conditions and different methods could lead to different results in detecting DIF. Therefore, using more than one DIF detection methods is also advised according to results of this study and current literature.

Limitations

There are limitations to mention about the study. The major limitation is that since the items were not released, it was not possible to identify sources of DIF by investigating the content. Identifying possible causes could give information to item developers to decrease the number of DIF items.

REFERENCES

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Arikan, S., van de Vijver, F., & Yagmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian Students in TIMSS 2007 and 2011. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*, 2039-2059. doi:10.12973/eurasia.2016.1268a

Atalay Kabasakal, K., Gok, B., Kelecioglu, H., & Arsan, N. (2012). Comparing different differential item functioning methods: A simulation study. *Hacettepe University Journal of Education*, *43*, 270-281.

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. doi:10.1037/1040-3590.6.4.284.

Cohen, J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mawhaw, NJ: Lawrence Erlbaum Associates.

Demir, S., & Köse, İ. A. (2014). An analysis of the differential item function through Mantel-Haenszel, SIBTEST and Logistic Regression Methods. *Journal of Human Sciences*, *11*(1), 700-714.

Doganay, A., & Bal, A. P. (2010). The measurement of students' achievement in teaching primary school fifth year mathematics classes. *Educational Sciences: Theory and Practice*, *10*, 199-215.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.

Gök, B., Kellecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*(156), 3-16.

Gün, F., & Baskan, G. A. (2014). New education system in Turkey (4+ 4+ 4): A critical outlook. *Procedia-Social and Behavioral Sciences*, *131*, 229-235.

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39-56). Charlotte, NC: Information Age Publishing.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (ETS Research Report No. RR-86-31). Princeton, NJ: ETS.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. En H.Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, N.J.: Erlbaum.

Jerrim, J. (2016). How Shift to Computer-based Tests Could Shake up PISA Education Rankings. *The Conversation*. Retrieved from http://theconversation.com/how-shift-to-computer-based-tests-could-shake-up-pisa-education-rankings-54869

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect'and what has been done about it?. *Oxford Review of Education*, *44*(4), 476-493.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Joldersma, K., & Bowen, D. (2010). *Application of Propensity Models in DIF Studies To Compensate For Unequal Ability Distributions*. Paper presented at the annual meeting of National Council on Measurement in Education, Denver, CO.

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381-399.

Komatsu, H., & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A view from East Asia. *Compare: A Journal of Comparative and International Education*, *47*(4), 616-623.

Liu, Y., Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. D. (2016). Investigating Causal DIF via Propensity Score Methods. *Practical Assessment, Research & Evaluation*, *21*(13), 1-24.

Lyons-Thomas, J., Sandilands, D. D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *Education &Science*, *39*(172), 20-32.

MEB (2015). *PISA 2012 Araştırması Ulusal Nihai Raporu*. Ankara. Retrieved from https://drive.google.com/file/d/0B2wxMX5xMcnhaGtnV2x6YWsyY2c/view

MEB (2016). *PISA 2015 Ulusal Raporu*. Ankara. Retrieved from http://pisa.meb.gov.tr/wp-content/uploads/2016/12/PISA2015_Ulusal_Rapor1.pdf

Muthen, B. O., & Muthen, L. K. (2015). *Mplus (Version 7.4)*. California. Los Angeles.

OECD (2016a). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Paris: OECD Publishing. doi:10.1787/9789264266490-en

OECD (2016b). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and FinancialLiteracy.* PISA, OECD Publishing, Paris.doi:10.1787/9789264255425-en

OECD (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.

Spaull, N. (2017). *Who Makes It Into PISA?: Understanding the Impact of PISA Sample Eligibility Using Turkey as a Case Study (PISA 2003 - PISA 2012).* OECD Education Working Papers, No. 154, OECD Publishing, Paris. http://dx.doi.org/10.1787/41d175fc-en

Temur, Ö. D. (2012). Analysis of prospective classroom teachers' teaching of mathematical modeling and problem solving. *Eurasia Journal of Mathematics, Science & Technology Education, 8*(2), 83-93. doi:10.12973/eurasia.2012.822a

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R. (2017). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (2nd, revised edition). New York, NY: Routledge.

Yıldırım, A., Özgürlük, B., Parlak, B., Gönen, E., & Polat, M. (2016). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. TC Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Yildirim, H. H., & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*(2), 108-121.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D. (1999). 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.

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

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1-23.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.

PISA 2015'de Türk Öğrencilerin Düşük Başarı Göstermelerinin Nedeni Değişen Madde Fonksiyonu (DMF) içeren maddeler midir?

GİRİŞ

Uluslararası Öğrenci Değerlendirme Programı (PISA) 15 yaşındaki öğrencilerin okuma, matematik ve fen okuryazarlığı alanlarındaki becerilerini uluslararası karşılaştırmalara olanak veren bir yapıda ölçmektedir. Katılan ülke sayısının giderek arttığı PISA'ya 70'in üzerinde ülke dahil olmaktadır. (OECD, 2016a). OECD üyesi olan Türkiye PISA'ya 2003 yılından beri düzenli olarak katılmaktadır. Ortalama puanın 500 olduğu PISA matematik okuryazarlık testinde, Türkiye PISA 2003'de 423, PISA 2006'da 424, PISA 2009'da 445, PISA 2012'de 448 ve PISA 2015'de 420 ortalama puan almıştır (MEB, 2015; MEB, 2016). Benzer bir değişim hem fen hem de okuma alanlarında da mevcuttur. PISA 2012'ye doğru artan yönde olumlu gelişmeler yaşanırken, 2015 yılında ciddi bir düşüşün yaşanması oldukça dikkat çekicidir. Bu düşüşün nedenlerinin araştırılması gerekmektedir. Nedenlerden bir tanesi ölçme aracında kullanılan maddelerin dil açısından yanlılık göstermeleri olabilir. Ölçme sonuçlarının sınavın uygulandığı dilden bağımsız olarak sonuçlar üretmesi beklenir. PISA soruları çoğunlukla İngilizce olarak geliştirilmekte, ardından diğer dillere adaptasyonu yapılmaktadır (OECD, 2017). Bu sebeple PISA matematik sorularının Türkçe ve İngilizce konuşan ülkelerdeki öğrenciler için değişen madde fonksiyonu (DMF) gösterip göstermediğinin incelenmesi gereklidir. Bu çalışmada Türkiye'deki öğrencilerin düşük puan alma nedeninin maddelerin DMF içermeleri olup olmadığı incelenecek, eğer neden bu değil ise de bu ihtimal elenerek, diğer ihtimallere odaklanılacaktır.

DMF tespit etme yöntemleri kullanılarak testlerin madde bazında yanlılık gösterip göstermediği ile ilgili ön inceleme yapılabilmektedir (Zumbo, 2007). DMF'nin ve sonrasında madde yanlılığının ortaya çıkması öğrenci gruplarının puanlarını doğru bir şekilde karşılaştırmayı engellemektedir. Aynı beceri düzeyine sahip iki öğrenci grubunun bir soruyu yanıtlama olasılıkları farklılaştığında DMF ortaya çıkmaktadır (van de Vijver & Leung, 1997; Zumbo, 2007). Bir maddede istatistiksel olarak DMF çıkarsa, uzmanlar o soruyu incelemeli ve neden DMF çıktığını yorumlayarak maddenin ilgili gruplar için yanlılık gösterip göstermediğine karar vermelidir (van de Vijver & Leung, 1997).

PISA soruları oldukça geniş bir uzman kadrosu tarafından titizlikle hazırlanmakta ve adaptasyon süreçleri gerçekleştirilmektedir (OECD, 2017). Ancak yine de, araştırmalar PISA matematik sorularında DMF içeren maddeler olduğunu raporlamışlardır (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009). Bu sebeple PISA 2015 maddelerini de DMF içerip içermedikleri bakımından incelemek faydalı olacaktır. Alan yazında PISA 2015 maddelerini Türk öğrenciler ve İngilizce konuşan öğrenciler bakımından DMF için karşılaştıran bir çalışmaya rastlanmamıştır.

Bu amaçla bu çalışmada Türk, İngiliz ve Amerikan öğrencilerin matematik sorularına verdikleri yanıtlar DMF içerip içermedikleri yönünden incelenmiştir. Türk öğrencilerin düşük matematik performansı gösterme nedenlerinden birisi olarak DMF içeren maddelerin olup olmaması incelenmiştir. Araştırma soruları ise

(1) Türk ve İngiliz öğrencileri karşılaştırıldığında, DMF içeren PISA 2015 matematik sorusu var mıdır?

(2) Türk ve American öğrencileri karşılaştırıldığında, DMF içeren PISA 2015 matematik sorusu var mıdır?

(3) DMF içeren maddeler testten çıkarıldığında matematik performans farklarından ortaya çıkan etki büyüklükleri değişmekte midir?

YÖNTEM

Örneklem

PISA 15 yaşındaki öğrencilerin ilgili konu alanlarındaki performanslarını ölçerken eksik test deseni kullanmaktadır (OECD, 2016b). Farklı kitapçıklar testin farklı sorularını içermektedir. Kitapçık 43, 45 ve 47 bir araya gelince tüm soruları içermektedir. Bu sebeple 43, 45, 47 numaralı kitapçıklara yanıt veren öğrenciler bu çalışmanın örneklemini oluşturmaktadır. Bu çalışmada 491 Türk, 1154 İngiliz ve 448 Amerikan öğrenci yer almaktadır.

Ölçme Aracı

PISA 2015 kapsamında öğrencilerin matematik performanslarının değerlendirmesi için toplam 69 madde kullanılmıştır. Her bir öğrenci yaklaşık 23 soru yanıtlamıştır. PISA matematik testindeki bu sorular ölçtükleri beceriler bakımından hiyerarşik bir yapıda hazırlanmıştır. En temel beceri olarak formüle etme, ardından uygulama ve en üst düzey düşünme süreci olarak yorumlama becerisi yer almaktadır (OECD, 2016b).

Veri Analizi

Bu çalışmada 3 farklı DMF belirleme yöntemi kullanılmıştır. Bu yöntemler logistik regresyon (LR), Mantel-Haenszel (MH) ve yapısal eşitlik modelidir (SEM). Her metot farklı hesaplama yöntemlerine dayalı olduğu için (Atalay Kabasakal, Gok, Kelecioglu & Arsan, 2012) daha tutarlı sonuçlar için en az 2 yöntemde farklılık gösteren maddeler DMF içeriyor olarak kabul edilmiştir. Logistik regresyon analizinde ilk adım olarak toplam puan, ikinci adım olarak toplam puan ve grup değişkeni, üçüncü adım olarak da toplam puan, grup değişkeni ve toplam puan ile grup değişkeninin etkileşimi modellere eklenmektedir. ΔR^2 0.035'den büyük ise DMF olduğuna karar verilmiştir (Jodoin and Gierl, 2001). SPSS programı kullanılarak bu analizler gerçekleştirilmiştir. Mantel-Haenszel metodunda ise grupların toplam puanına göre K adet 2x2 çapraz tablolar baz alınarak ki-kare değerleri hesaplanmaktadır. Daha sonra ilgili dönüşümler yapılarak MH D-DIF indeksi oluşturulmaktadır (Holland & Thayer, 1986). Bu değer 1'den büyük ise DMF olduğuna karar verilmektedir (Zieky, 1993). DIFAS 5.0 programı ile hesaplamalar yapılmıştır (Penfield, 2005). SEM ile DMF belirleme yönteminde ise doğrulayıcı faktör analizinde ilgili parametrelerin eşit olmaya zorlanması sonucunda elde edilen fit değerlerine büyük etkisi olan maddeler DMF içeren madde olarak belirlenmektedir (van de Vijver, 2017). Comparative fit index (CFI) ve Tucker Lewis index (TLI) değerleri arasındaki fark 0.010'dan büyük ise modifikasyon indeksleri incelenerek DMF içeren maddeler tespit edilir (Cheung and Rensvold, 2002). Bu analizde Mplus 7.4 programı kullanılmıştır (Muthen & Muthen, 2015).

SONUÇ VE TARTIŞMA

İç Tutarlılık

PISA 2015 matematik sınavı için Cronbach's alpha iç tutarlılık katsayıları kitapçık 43, 45 ve 47 için Türk öğrenciler için sırasıyla 0.78, 0.79, 0.76; İngiliz öğrenciler için 0.81, 0.84, 0.85; ve Amerikan öğrenciler için 0.80, 0.86, 0.86 olarak hesaplanmıştır. Bu değerler testin iyi düzeyde iç tutarlılığa sahip olduğunu göstermektedir (Cicchetti, 1994).

DMF sonuçları

Bu kısımda LR, MH ve SEM yöntemleri kullanılarak elde edilen DMF sonuçları verilmektedir.

LR yöntemi ile elde edilen sonuçlar Tablo 4'de verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_8, B47_9 ve B47_19), Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 14 tanesi (B43_11, B43_15, B43_16, B45_10, B45_11, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_11, B47_14 ve B47_19) DMF içermektedir. MH yöntemi ile elde edilen sonuçlar Tablo 5'de verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_10 ve B47_19) Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_7, B47_9, B47_11, B47_14 ve B47_19) DMF içermektedir. SEM yöntemi ile elde edilen sonuçlar Tablo 6'da verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 4 tanesi (B45_2, B45_10, B45_13, B45_18) Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 2 tanesi (B45_13 ve B47_9) DMF içermektedir.

En az iki yöntem tarafından DMF içerdiği görülen maddeler burada listelenmiştir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 9 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9 ve B47_19) her iki yönteme göre DMF içermektedir. Ayrıca, hangi maddelerin hangi grubun lehine çalıştığının raporlanması da önem taşımaktadır. Bu 9 maddeden 3 tanesi Türk öğrenciler lehine (B43_11, B45_10, B45_18, B47_6, B47_7, B47_9) 3 madde ise İngiliz öğrencilerin lehine çalışmaktadır (B45_13, B47_1, B47_19). Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_11, B45_13, B47_1, B47_7, B47_9, B47_11, B47_14 ve B47_19) her iki yönteme göre DMF içermektedir. Bu 10 maddeden 5 tanesi Türk öğrenciler lehine (B43_11, B45_10, B47_7, B47_9, B47_14) 4 madde ise Amerikan öğrencilerin lehine çalışmaktadır (B45_13, B47_1, B47_11, B47_19). Bir madde (B45_11) kısmen Türk öğrencilerin lehine, kısmen ise Amerikan öğrencilerin lehine çalışmaktadır. Türk-İngiliz ve Türk-Amerikan karşılaştırmaları benzer sonuçlar vermiştir.

Tablo 8 incelendiğinde, DMF gösteren tüm maddelerin açık uçlu sorular olduğu görülmektedir. Ayrıca, Türk öğrencilere hem İngiliz hem de Amerikalı öğrencilere göre avantaj sağlayan 7 sorunun 4 tanesinin en alt düşünme sürecini ölçen formüle etme düşünme süreci ile ilgili olduğu görülmektedir. Formüle etme becerisini ölçen hiçbir soru İngiliz ve Amerikan öğrencilerin lehine çalışmamaktadır.

DMF Sonuçları ve Etki Büyüklüğü

Türk öğrenciler ile İngiliz ve Amerikalı öğrenciler arasında başarı farkı bulunmaktadır. Gruplar arası farkları örneklemdeki kişi sayısından bağımsız olarak değerlendirebilmek için etki büyüklüğünü kullanmak iyi bir yöntemdir (Field, 2013). Tablo 9'da öğrenci grupları arasındaki farkın etki büyüklüğü tüm maddeler kullanılarak ve DMF gösteren maddeler çıkarıldığında hesaplanmıştır. Türk ve İngiliz öğrenciler arasında başlangıçta .51 ile .93 arasında değişen etki büyüklüğü hesaplanmıştır. DMF içeren maddeler çıkarıldığında ise bir değişiklik gözlenmemiştir. Aynı şekilde Türk ve Amerikalı öğrenciler arasında .28 ile .85 arasında değişen etki büyüklüğü gözlenmiştir. DMF içeren maddeler çıkarıldığında yine farkın değişmediği görülmüştür.

Tartışma

Bu çalışma Türk öğrencilerin PISA 2015 matematik testinden çok düşük alma nedenlerinden birisi olabilecek olan DMF içeren maddeleri incelemesi bakımından oldukça önemlidir. Araştırmada önceki bölümlerde belirtildiği gibi DMF içeren maddeler tespit edilmiştir. Ancak, bu maddeler sadece Türk öğrencilerin aleyhinde çalışmamaktadır. DMF içeren maddelerin bir kısmı Türk öğrencilerin lehine çalışmaktadır. Ek olarak, etki büyüklükleri karşılaştırıldığında DMF içeren maddelerin toplam puanlarda herhangi bir gruba bir avantaj sağladığına dair kanıt bulunmamaktadır.

Puanlardaki düşüş için farklı nedenlere odaklanmak gerekmektedir. Türk öğrencilerin PISA 2015 ortalama matematik puanlarında neden düşüş yaşadıklarını tespit etmek için yıllar içerisinde seçilen örneklemlerin karşılaştırılabilirliği, sınavın kağıt kalem formatı yerine artık bilgisayar ortamında uygulanması ve ülke bazındaki eğitim sistemi, öğretim programları ve eğitim politikalarında yaşanan değişimler gibi farklı değişkenleri de incelemek gerekmektedir.

PISA'daki sorular yayınlanmadığı için DMF içeren maddelerin yanlılık gösterip göstermediğine dair uzman incelemesi yaptırılamamıştır. Ancak, soruların özellikleri incelendiğinde bazı önemli ipuçları elde edilmiştir. DMF içeren tüm maddelerin açık uçlu sorulardan oluşması bu soruların puanlanma süreçlerinin yeniden gözden geçirilmesi gerektiğini göstermektedir. Bu puanlama sırasında maddeler DMF içeriyor hale gelmiş olabilir. Diğer bir bulgu da, Türk öğrencilerin lehine çalışan maddelerin çoğunun en alt düzey düşünme sürecini içeren maddeler olmasıdır. Bu tip maddelerin hiçbiri İngilizce konuşan öğrencilere DMF göstermemiştir. Türkiye'deki eğitim genel olarak çok soru çözmeye dayandığı için, öğrenciler temel becerileri geliştirmiş ve bu tip sorularla daha fazla karşılaşmış olabilir (Arikan, van de Vijver & Yagmur, 2016; Doganay & Bal, 2010; Temur, 2012). Bu durum da bu tip maddelerin Türk öğrenciler lehine DMF göstermiş olabileceği anlamına gelmektedir. Son olarak, kullanılan DMF belirleme yöntemleri karşılaştırıldığında logistik regresyon ve Mantel-Haenszel yöntemlerinin yapısal eşitlik modeline göre birbirine daha yakın sonuçlar verdiği görülmüştür.

Appendix A.

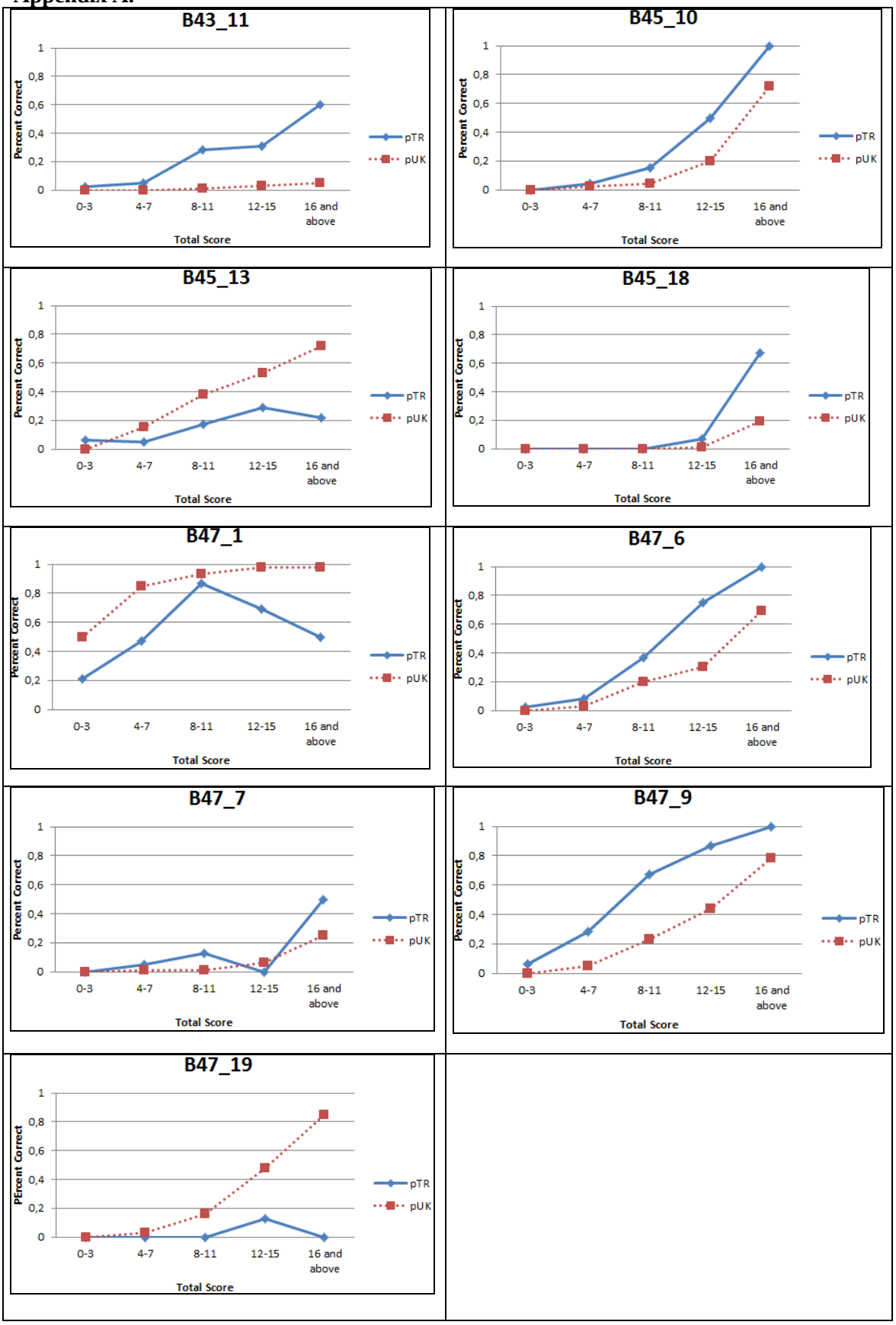

Figure 1. Graphical Representation of DIF Items for TR and UK Students

Appendix B.

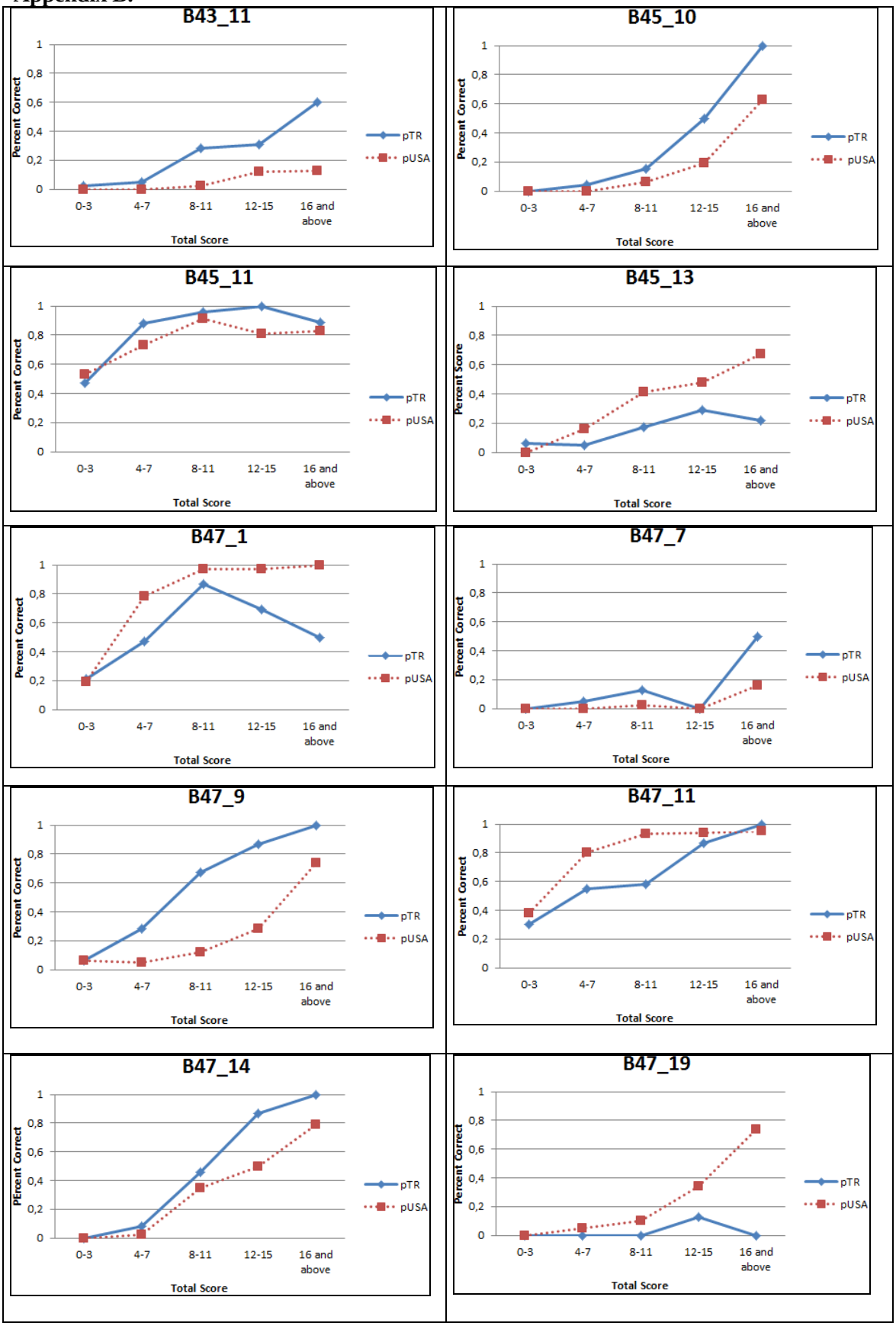

Figure 2. Graphical Representation of DIF Items for TR and USA Students