IJETS@R

(ISSN: 2587-0238)

Demirus, K. B. & Pektaş, S. (2022). Investigation of TIMSS 2015 science test items in terms of differential item functioning according to language and culture. *International Journal of Education Technology and Scientific Researches, 7(18), 1166-1178.*

DOI: http://dx.doi.org/10.35826/ijetsar.499

Article Type: Research Article

INVESTIGATION OF TIMSS 2015 SCIENCE TEST ITEMS IN TERMS OF DIFFERENTIAL ITEM FUNCTIONING ACCORDING TO LANGUAGE AND CULTURE*

Kadriye Belgin DEMİRUS
Asst. Prof. Dr., Başkent University, Ankara, Turkey, belgindemirus@gmail.com
ORCID: 0000-0003-2570-3653

Sami PEKTAŞ
Asst. Prof. Dr., Niğde Ömer Halisdemir University, Niğde, Turkey, pektassami@gmail.com
ORCID: 0000-0003-4753-6112

Received: 02.03.2022 Accepted: 18.05.2022 Published: 15.06.2022

ABSTRACT

In this study, it was investigated whether the science test items in the Trends in International Mathematics and Science Study (TIMSS) 2015 exam included differential item functioning (DIF) in terms of culture and language. DIF analyses were performed using the Mantel Haenszel (MH) method in the RStudio program. In the research, Turkey, Australia, New Zealand, Morocco and Egypt countries were studied. For the countries in the study group, data analysis of 16 multiple-choice items of the science achievement test in the 8th grade 10th booklet of TIMSS 2015 was carried out. In linguistic and cultural comparisons of countries, the fact that they are in the same and different success groups has been taken into account. As a result, it was observed that the order of achievement was not effective in the number of items with DIF. It was concluded that the number of items with moderate and high level DIF was the highest among countries with different cultures and speaking different languages. It was observed that the number of items with DIF was low among the countries speaking the same language and in different cultures. It was concluded that the differentiation of the spoken language was a more effective variable in determining the item with DIF. Therefore, based on the findings of the research, it is suggested that the translation and adaptation processes should be carried out meticulously in international tests such as TIMSS and PISA. Future DIF examinations in terms of culture and language can be carried out by selecting different countries.

Keywords: Differential item functioning, Mantel Haenszel, science achievement, TIMSS, language and culture.

* This study was presented as an oral presentation at the 6th International Congress on Measurement and Evaluation in Education and Psychology.

INTRODUCTION

In our country and in the world, national exams such as Student Achievement Determination Exam (SADE) and international exams such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are applied in order to determine student achievement. Important decisions are taken about individuals in national exams and about countries in international exams. In order for the decisions taken to be valid and reliable, the validity and reliability of the measurements must be high. In these exams, apart from variables such as gender, place of residence, family education level, etc., characteristics such as language, culture and race also differ among students who take the exam. Different characteristics of individuals cause the results obtained from the measurements to be different (Asil & Gelbal, 2012; Ercikan, 1998; Ercikan, 2002; Toprak & Yakar, 2017; Uzun Başusta, 2010; Uzun & Gelbal, 2017). However, it is not correct to interpret this difference as the result of only individual characteristics. The reason may be that this difference between individuals is due to the characteristics of the measurement tool (Uzun Başusta, 2010). Individuals with similar abilities but in different groups should not find test items easier or more difficult. In short, the correct answering behaviors of groups with equal ability level should not be affected by the group they are in. This causes differences in test scores of individuals. Therefore, when individuals want to be ranked according to a certain ability, unfair and wrong decisions can be made. For example, individuals who take the test in a certain culture or language group should not answer the question more easily because of their group. This concept, called item bias, is a feature that originates from the measurement tool and reduces its validity. Item bias is a concept that needs to be taken care of because it causes systematic errors that distort the measurement results in selecting and ranking students (Zumbo, 1999).

The process of investigating item bias includes both the statistical analysis of the items and the examination of the source of the differences between these items. If the difference in the probability of answering the item correctly by groups of equal ability level (different groups in terms of culture, gender, socio-economic level, language, etc.) is only due to statistical aspects, it is called “differential item functioning (DIF)” (Hambleton at all, 1991; Osterlind, 1983; Raju, 1990; Zumbo, 1999). DIF must not use as synonymous with bias, ignoring the reasoning behind the name change. DIF analysis indicates that there is a difference in the subgroups and this difference is significant, but does not indicate which variable is the cause of this difference (Osterlind & Everson, 2009). The fact that an item shows DIF as a result of statistical analyses does not indicate that that item is biased in favor of a group, because the item with DIF may also indicate a real difference between the groups in terms of measured ability. Items showing DIF are reviewed by taking expert opinion. It is decided whether the items are biased or not and what the source of measurement errors is. No test measures perfectly an intented trait or knowledge domain, but as long as measurement error affects scores for members of different groups equally, a test is not biased (Camilli & Shepard, 1994; Zumbo, 2007). For example, wrong translations into the target language, translations that are not suitable for the target culture will cause erroneous results due to the measurement tool. There are various studies examining the effect of culture and language differences on measurement results in international tests. The majority of these studies are

examining the PISA exam which is focused on scientific literacy (Asil & Gelbal, 2012; Cheema, 2019; Çıkrıkçı Demirtaşlı & Ulutaş, 2015; Demir & Köse, 2014; Gök, Atalay Kabasakal & Kelecioğlu, 2014; Gür, 2019; Güzeller, 2011; Kıbrıslıoğlu, 2015; Köse, 2015; Sırgancı & Çakan, 2020). When the current literature was examined, it was observed that a few studies were held on TIMSS. Karakoc, et. all (2016) aimed to examine the measurement invariance of TIMSS 2011 mathematics test in terms of different cultures, but not DIF. This situation can be raised the importance of the research.

In a study conducted by Ercikan (2002), it was examined whether the TIMSS 1995 English and French versions of the science and mathematics items had differential item functioning. It was concluded that the items with DIF were affected by the problems in adaptation, the difference in familiarity with the item formats, the difference in the curriculum, and the cultural difference. Therefore, the items were also found to be biased. Methodologically, the intent of internal item bias analyses was to differentiate between true group differences and bias in the measurement. Also group differences on test items could not be expressed automatically as evidence of bias; because in reality, score differences might be valid reflections of groups differences in knowledge and experience. But, the less the situation of being affected by these differences, the higher the validity of the measurement results will be. Bias is a validity problem in the measurement literature. It means that there is a systematic error in the measurement results in favor of or against a certain group. Holland & Thayer (1988) refers to the uninterpreted relative difficulty as differential item functioning or DIF. But it is held that the process of using such an index to detect bias in conjunction with logical analysis as an item bias detection procedure. By the way test developers should never refer on the basis of item bias procedures alone that a test has been guaranteed free of bias and valid for all possible use (Camilli ve Shepard, 1994). The bias research was excluded from the scope of this study.

TIMSS research, conducted by the International Association for the Evaluation of Educational Achievement (IEA), describes achievement in four international criteria across the scale of scientific achievement for evaluating fourth- and eighth-grade students in participating countries: Advanced, Upper, Moderate and Low Level. There are clear challenges in advancing students in science achievement around the world. In terms of the percentage of students who met the criteria, countries were able to bring 7% of their eighth-grade students to advanced science achievement and 29% to upper science achievement (IEA, 2016). Science achievement of secondary school students in Turkey has been increasing compared to previous years. The majority of Turkish students (59%) were moderate (31%) and upper level (28%) in international science proficiency. Over the years, there has been a decrease in the percentage of low and below-low level students, while the proportion of advanced, upper and moderate students has increased. However, in recent years, there has been a significant increase in Turkish students' positive attitudes towards science teaching (interest, love, self-esteem and value). However, teachers' participation in science-related professional activities (identifying students' needs, using information technologies, teaching science, developing skills, etc.) is far behind the world average (MEB, 2016). In this study, countries with moderate and below-low levels of science achievement were studied.

TIMSS was first applied to fourth and eighth grade students in the world in 1995. Turkey did not participate in the 1995 and 2003 studies. Turkey participated in 1999 and 2007 studies only at the eighth-grade level, and in 2011 and 2015 studies at the fourth and eighth grade level (MEB, 2016). TIMSS 2015 was attended by more than 580,000 students from 57 countries and in total. Most of the TIMSS items evaluate students' practice and reasoning skills (Murtin, Mullis, Foy & Hooper, 2015). It is an important issue whether these evaluations are valid in cultural and linguistic terms. In this context, "Does 8th grade science test items in TIMSS 2015 exam include differential item functioning in terms of culture and language?" constitutes the research question. Within the framework of the research question, answers to the following questions were sought.

1) Is there any item with DIF in the comparison of Turkey and Australia in the TIMSS 2015 science test?
2) Is there an item with DIF in the comparison of Turkey and Egypt in the TIMSS 2015 science test?
3) Is there an item with DIF in the comparison of Egypt and Morocco in the TIMSS 2015 science test?
4) Is there an item with DIF in the comparison of Australia and New Zealand in the TIMSS 2015 science test?

METHOD

The Research Model

In this study, it was investigated whether the science test items in the TIMSS 2015 exam included differential item functioning in terms of culture and language, by making cross-country comparisons. The research is descriptive research that reveals the current situation. In descriptive research, it is tried to explain the relationships between the variables examined (Brown, Cozby, Kee & Worden, 1999).

Population-Sample

The population of the study consists of 46 countries participating in TIMSS 2015 at the eighth-grade level. The sample of the research consists of eighth grade students from five countries (Turkey, Australia, Egypt, Morocco and New Zealand) selected by purposive sampling method from these countries. The language that reflects the intercultural difference has been effective in the selection of the countries. Countries with different languages and cultures were compared according to their status of having the same achievement level (Turkey-Australia) and different achievement levels (Turkey-Egypt). Also, a comparison of [Egypt (Arabic)-Moroccan (Arabic)] countries that took the test in the same language but are from different culture was made. In addition, in order to reveal the relative effects of language and culture, countries that took the test in the same language and whose cultures were the same (Australia-New Zealand) and different (Egypt-Morocco) were also included in the study. Table 1 shows the frequency distribution of 8th grade students in the TIMSS sample according to country.

Table 1. Frequency Distribution of TIMSS 2015 Samples According to Country

Countries	Frequency (f)	Percent Frequency (%f)
Turkey	6079	13.0
Australia	10338	23.0
Egypt	7822	17.0
Morocco	13035	29.0
New Zealand	8142	18.0
Total	45416	100

As seen in Table 1, a total of 45416 students were included in the 8th grade sample of TIMSS. The IEA (2016) did not share the results of all of these students. Within the scope of the research, students who filled out the 10th booklet of all countries formed the sample of the study. The frequency distribution of the sample of the study is shown in Table 2.

Table 2. Frequency Distribution of The Students Forming The Sample of Study According to Country

Countries	Frequency (f)	Percent Frequency (%f)
Turkey	423	13.0
Australia	751	23.0
Egypt	567	18.0
Morocco	942	29.0
New Zealand	531	17.0
Total	3214	100.0

When Table 2 is examined, the sample of Turkey with 423 students constitutes 13% of the study group, Australia 23% with 751 students, Egypt 18% with 567 students, Morocco 29% with 942 students and New Zealand 17% with 531 students. It is seen that the number of students that can be reached is 3214 in 5 countries.

Data Collection Tool

The data collection tool of the research is the TIMSS 2015 science multiple choice achievement test conducted by the IEA. The data required for the study were obtained from the IEA, TIMSS website (https://timss.bc.edu/timss2015/international-database/). Science learning domains consist of biology, physics, chemistry and earth sciences. In addition, items belonging to each learning domain were written for three cognitive levels (knowing, applying, reasoning). Table 3 below shows the number of multiple-choice items pertaining to TIMSS 2015 science learning domains and cognitive levels.

Table 3. Frequency Distribution of TIMSS 2015 Science Questions According to Learning Domain and Cognitive Levels

Science Learning Domains	Cognitive Level			Total Frequency	% Total Frequency
	Knowing	Applying	Reasoning		
Biology	18	14	3	35	31.8
Physics	10	13	8	31	28.2
Chemistry	14	4	1	19	17.3
Earth Sciences	16	8	1	25	22.7
Total	58	39	13	110	100

TIMSS 2015 science test consists of 14 booklets. In the investigation of the research problem, the questions on the cognitive levels of knowing, applying and reasoning in the learning domains of biology, physics, chemistry and earth sciences belonging to the same booklet (booklet 10) were discussed. The frequency distribution of the items constituting the data collection tool according to learning domain and cognitive levels is shown in Table 4.

Table 4. Frequency Distribution of The Science Questions of The Data Collection Tool According to Learning Domain and Cognitive Levels

Science Learning Domains	Cognitive Level			Total Frequency	% Total Frequency
	Knowing	Applying	Reasoning		
Biology	4	1	-	5	31.3
Physics	1	4	1	6	37.5
Chemistry	2	-	1	3	18.7
Earth Sciences	1	-	1	2	12.5
Total	8	5	3	16	100

As seen in Table 4, the data collection tool of the research is the science multiple choice test consisting of 16 questions in total at the cognitive levels of 8 knowing, 5 applying and 3 reasoning.

Data Analysis

The determination of the items showing DIF was carried out using the Mantel-Haenzel (1959) method in the RStudio program. R is an open-source program that is distributed free of charge over the internet and can run on almost all operating systems. R libraries are developed with ready-made code and functions and provide convenience to users (Team, 2013). In DIF determination methods, the "difR" package was loaded and commands were written for analysis. The "difR" library was developed with the aim of examining two-category data in terms of DIF with various methods. Through the codes in this library, uniform and non-uniform DIF can be examined based on KTK and MTK (Magis et al., 2016).

DIF determination method based on the MH method, taking into account the ΔMH value, there are three classifications reflecting the DIF level accepted by the ETS (Educational Testing Service). This classification system, organized by Zieky (1993), is given in Table 5 (Camilli & Shepard, 1994. p.121).

Table 5. DMF Classification System

Value Range	DIF Level	Explanation
IΔMHI <1	A	DIF is low or negligible
1 ≤ IΔMH I <1.5	B	DIF is moderate
IΔMH I≥1.5	C	DIF is high

A negative Δ_{MH} value is interpreted as showing DIF in favor of the individuals in the reference group, being positive as showing DIF in favor of the individuals in the focus group, and being equal to zero as not having DIF (De Ayala, 2009). Within the scope of the research, before starting the DIF process, confirmatory factor analysis (DFA) regarding whether 16 items in the 10th booklet were collected in a single latent variable or not was

tested with the IBM AMOS-23 program for 5 countries and it was concluded that the fit index values showed a perfect fit. The results of confirmatory factor analysis are given in Table 6.

Table 6. Fit Index Values For Confirmatory Factor Analysis

Countries	CMIN/DF	RMSEA	GFI	AGFI	CFI	TLI
Turkey	1.149	0.019	0.97	0.96	0.95	0.95
Australia	1.163	0.015	0.98	0.97	0.96	0.96
Egypt	1.081	0.012	0.98	0.97	0.96	0.95
Morocco	0.995	0.000	0.99	0.98	1.00	1.00
New Zealand	1.14	0.016	0.97	0.96	0.97	0.96

Looking at Table 6, it is seen that Morocco provides the best fit in the first-level single-factor (16-item) CFA model regarding science achievement among 5 countries.

FINDINGS

In this section, it is discussed whether the 8th grade science test items in the TIMSS 2015 exam show differential item functioning in terms of culture and language. The findings of the MH test used for this purpose are given within the scope of the research questions.

1) Findings related to the question "Is there an item with DIF in the comparison of Turkey and Australia in the TIMSS 2015 science test?" are shown in Table 7.

Table 7. Analysis of Science Items Based On MH According to Turkey and Australia Countries

Item No.	α-MH	$\Delta - MH$	DIF Class	DIF Type	Group Showing DIF In Their Favor
I1	0.6671	0.9514	A		
I2	1.853	-1.4495	B	Moderate	Australia
I3	1.5895	-1.089	B	Moderate	Australia
I4	2.2611	-1.9173	C	High	Australia
I5	2.0524	-1.6896	C	High	Australia
I6	1.2995	-0.6156	A	High	Australia
I7	8.8383	-5.1209	C	High	Australia
I8	2.3638	-2.0216	C	High	Australia
I9	1.9293	-1.5444	C	High	Australia
I10	2.7636	-2.3888	C	High	Australia
I11	0.4148	2.0681	C	High	Australia
I12	1.4448	-0.8647	A	High	Australia
I13	2.2775	-1.9342	C	High	Australia
I14	2.4698	-2.1247	C	High	Australia
I15	0.0564	6.7584	C	High	Turkey
I16	2.2736	-1.9302	C	High	Australia

When Table 7 is examined, 2 science items out of 16 show moderate DIF and 11 science items show high DIF. Except for 15th item, which showed a high level of DIF, all DIF items were found to have DIF in favor of Australia. Only 15th item showed DIF in favor of Turkey.

2) Findings related to the question "Is there an item with DIF in the comparison of Turkey and Egypt in the TIMSS 2015 science test?" are shown in Table 8.

Table 8. Analysis of Science Items Based On MH According to Turkey and Egypt Countries

Item No.	α-MH	$\Delta - MH$	DIF Class	DIF Type	Group Showing DIF In Their Favor
I1	2.4489	-2.1048	C	High	Egypt
I2	1.2199	-0.4672	A		
I3	5.7943	-4.1287	C	High	Egypt
I4	1.4677	-0.9017	A		
I5	1.7011	-1.2484	B	Moderate	Egypt
I6	2.5055	-2.1585	C	High	Egypt
I7	2.5916	-2.2378	C	High	Egypt
I8	2.2819	-1.9388	C	High	Egypt
I9	4.9199	-3.7442	C	High	Egypt
I10	1.4493	-0.872	A	High	
I11	0.1587	4.3265	C	High	Turkey
I12	3.737	-3.098	C	High	Egypt
I13	6.1911	-4.2843	C	High	Egypt
I14	1.8234	-1.4116	B	Moderate	Egypt
I15	0.0328	8.0281	C	High	Turkey
I16	1.4651	-0.8975	A		

According to Table 8, 2 of the 16 items showed moderate DIF and 10 science items showed high DIF. Except for 11th and 15th items, which show a high level of DIF, all DIF items were found to be DIF in favor of Egypt. 11th and 15th items in the science test showed DIF in favor of Turkey.

3) Findings related to the question "Is there an item with DIF in the comparison of Egypt and Morocco in the TIMSS 2015 science test?" are shown in Table 9.

Table 9. Analysis of Science Items Based On MH According to Egypt and Morocco Countries

Item No.	α-MH	$\Delta - MH$	DIF Class	DIF Type	Group Showing DIF In Their Favor
I1	0.303	2.8058	C	High	Egypt
I2	1.5354	-1.0076	B	Moderate	Morocco
I3	1.2408	-0.5071	A		
I4	1.1411	-0.3102	A		
I5	0.8138	0.4843	A		
I6	1.4233	-0.8295	A		
I7	3.4236	-2.8921	C	High	Morocco
I8	1.3869	-0.7686	A		
I9	1.5087	-0.9665	A		
I10	0.7287	0.7436	A		
I11	0.7708	0.6117	A		
I12	0.5643	1.3448	B	Moderate	Egypt
I13	0.5709	1.3175	B	Moderate	Egypt
I14	1.1127	-0.251	A		
I15	0.9542	0.1103	A		
I16	0.4323	1.9707	C	High	Egypt

When Table 9 was examined, it was observed that 3 science items out of 16 items had a high level of DIF and two showed DIF in favor of Egypt. In the science test, it was seen that 3 items showed moderate DIF, 2 of which showed DIF in favor of Egypt and 1 of which showed DIF in favor of Morocco.

4) Findings related to the question "Is there an item with DIF in the comparison of Australia and New Zealand in the TIMSS 2015 science test?" are shown in Table 10.

Table 10. Analysis of Science Items Based On MH According to Australia and Zealand Countries

Item No.	α-MH	$\Delta - MH$	DIF Class	DIF Type	Group Showing DIF In Their Favor
I1	1.493	-0.9419	A		
I2	0.9265	0.1794	A		
I3	1.0807	-0.1824	A		
I4	0.7713	0.6103	A		
I5	1.0614	-0.1401	A		
I6	0.9006	0.2461	A		
I7	0.9206	0.1943	A		
I8	1.0225	-0.0523	A		
I9	0.613	1.1499	B	Moderate	New Zealand
I10	1.0842	-0.1901	A		
I11	1.0279	-0.0647	A		
I12	1.212	-0.4518	A		
I13	1.0511	-0.1171	A		
I14	1.4789	-0.9195	A		
I15	0.9003	0.2468	A		
I16	1.2931	-0.6041	A		

According to Table 10, only the 9th item among 16 science items was found to have a moderate DIF in favor of New Zealand.

5) The status of containing DIF in terms of culture and language of the 8th grade science test items in the TIMSS 2015 exam is summarized in Table 11 in line with the research findings:

Table 11. Frequency Distribution of Items with DIF In Terms of Culture and Language

Countries	Achievement Levels	Total Item Frequency (f_T)	Item with DIF Frequency (f)	Percent Frequency (%f)
Turkey-Australia (Different Culture - Different Language)	Same (Moderate)	16	13	81.25
Turkey – Egypt (Different Culture - Different Language)	Different (Moderate- Below Low Level)	16	12	75
Egypt-Morocco (Different Culture - Same Language)	Same (Below Low Level)	16	6	37.5
Australia-New Zealand (Same Culture - Same Language)	Same (Moderate)	16	1	6.25

Looking at Table 11, 13 (81.25%) out of 16 items show DIF in the comparison of Turkey and Australia, which have moderate achievement according to TIMSS 2015 exam results, and also in the comparison of countries with different cultures and different languages. According to the TIMSS 2015 exam results, in the comparison of Turkey, which has a medium level of achievement, and Egypt, which has a achievement below-low level, 12 out of 16 items (75%) show DIF in the comparison of countries with different cultures and different languages. According to the TIMSS 2015 exam results, 6 out of 16 items (37.5%) show DIF in the comparison of Egypt and Morocco, which have below-low level achievement, and also in the comparison of different cultures and the same language. According to the results of TIMSS 2015, 1 (6.25%) of 16 items shows DIF in the comparison between Australia and New Zealand, which has moderate achievement, and also in the comparison of the same culture and same language.

CONCLUSION and DISCUSSION

When the findings are examined, it is observed that the number of items with DIF increases significantly when there are differences in both culture and language between countries. However, the lowest number of items with DIF was obtained among countries with the same culture, language and achievement levels (Australia-New Zealand).Similarly, Asil& Gelbal (2012), Atalay, Kabasakal& Kelecioğlu (2012), Gür (2019), Köse (2015), Uzun & Gelbal (2017) and Sırgancı & Çakan (2020) found in their studies that as language and culture differences between groups increase, number of items showing DIF increases. Especially, in the study that was conducted by Sırgancı& Çakan (2020) put forth that all of PISA 2006 student questionnaire items showed DIF when both cultural and linguistic differences occured between Australia and Turkey. Therefore, the number of items with DIF may have increased due to the problems experienced in the translation and adaptation processes into different languages. Parallel to this finding, in the study by Asil & Gelbal (2012), it was stated that problems arising from translation and adaptation affected the number of items with DIF. The fact that the least number of DIF items in the study is seen among countries with the same culture and same language (New Zealand and Australia) shows parallelism with the research findings of Gök, Atalay Kabasakal & Kelecioğlu (2014).

It was also observed that the number of items with DIF decreased somewhat when the achievement levels differed in countries with different languages and cultures. While 81.25% of the items showed DIF in countries with different languages and cultures at the same level of achievement (Turkey-Australia), this rate decreased to 75% in countries with different languages and cultures with different levels of achievement (Turkey-Egypt). There was only one item difference in the number of items. This result supports the finding of Gök, Atalay Kabasakal & Kelecioğlu (2014) and Uyar & Kaya Uyanık (2016) in their studies that “the results are not affected by the achievement order of the countries”. Although there are different cultures and languages in Turkey and Egypt, it may have been caused by the fact that it is closer to Egypt when compared to Australia in terms of geography. Thus, the determination power of DIF, its number, the realized error are affected by many variables such as sample size and DIF amount, as well as the achievement level of the groups compared (Erdem Keklik, 2014). In another finding, while 37.5% of the items showed DIF in countries with different cultures and the same language with the same achievement level, this rate decreased to 6.25% in countries with the same culture and language with the same achievement level. This finding can be interpreted as the cultural difference is effective in the number of DIFs when the level of achievement and language are the same.

The mediation of language is indispensable for the development of language and for the acceptance and spread of cultural elements. Culture and language are inseperable parts of a nation (Göçer, 2013). Language shapes culture; but alone is not enough. Culture is also affected by different phenomena besides language. Culture consists of four kinds of symbols: values, norms, beliefs and finally, expressive symbols (Peterson, 1979). Thus, in this study, countries where the language is the same but the culture is different were compared with each other. Similar to the research findings, in studies which investigate only cultural differences show

that the differentiation of cultures increases the number of items with DIF (Demir& Köse, 2014; Uyar & Kaya Uyanık, 2016).

RECOMMENDATIONS

Items with DIF to be arised in exams that determine the strengths and weaknesses of education policies at the international level, develop new policies, and compare the knowledge and skills of students with the world's countries have a concern for bias. Accordingly, evaluating and interpreting the results of a low-valid test and developing education policies can lead to wrong decisions. Therefore, based on the findings of the research, it is suggested that the translation and adaptation processes should be carried out meticulously in international tests such as TIMSS and PISA.

Future DIF examinations in terms of culture and language can be carried out by selecting different countries. Items showing DIF can be examined depending on expert opinions whether they are biased or not. Different tests of international exams such as TIMMS, PISA and PIRLS and DIF evaluation of survey items can be tested with different DIF methods.

ETHICAL TEXT

In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. Responsibility for any violations that may arise regarding the article belongs to the author.

Author(s) Contribution Rate: In this study, the contribution rate of the first author is 50% and the contribution rate of the second author is 50%.

REFERENCES

Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. *Education and Science, 37*(166), 236-249.

Atalay Kabasakal, K. & Kelecioğlu, H. (2012). Evaluation of attitude items in PISA 2006 student questionnaire in terms of differential item functioning. *Ankara University, Journal of Faculty of Educational Sciences, 45*(2), 77-96. https://doi.org/10.1501/Egifak_0000001254

Brown, W.K., Cozby, C.P., Kee, D.W., & Worden, P.E. (1999). *Research methods in human development* (2.b.). Mayfield.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items.* Sage.

Cheema, J. R. (2019). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology, 16*(2), 152-166. https://doi.org/10.1080/17405629.2017.1358607

Çıkrıkçı Demirtaşlı N., & Ulutaş, 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

De Ayala, R.J. (2009). *Methodology in the social sciences. The theory and practice of item response theory.* Guilford.

Demir, S., & Köse, İ. (2014). An analysis of the differential item function through Mantel-Haenszel, SIBTEST and logistic regression methods. *International Journal of Human Sciences*, *11*(1), 700-714. doi:10.14687/ijhs.v11i1.2798

Ercikan, K. (1998). Translation effects in international assessments, *International Journal of Educational Research*, 29(6), 543-553. https://doi.org/10.1016/S0883-0355(98)00047-0

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(34), 199-215. https://doi.org/10.1080/15305058.2002.9669493

Erdem Keklik, D. (2014). Comparison of Mantel-Haenszel and logistic regression techniques in detecting differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 12-25. https://doi.org/10.21031/epod.71099

Göçer, A. (2013). The opinion of Turkish student teachers on the relationship between language and culture: a phenomenological analysis. *Erzincan University: Journal of Faculty of Education*, *15*(2). 25-38

Gök, B., Atalay Kabasakal K., & Kelecioğlu, H. (2014). Analysis of attitude ıtems ın PISA 2009 student questionnaire ın terms of differential ıtem functioning based on culture. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(1), 72-87. https://doi.org/10.21031/epod.64124

Gür, E. (2019). *An Investigation of The PISA 2015 in terms of differantıal ıtem functioning based on culture* (Unpublished Master Thesis). Hacettepe University: Institute of Educational Sciences.

Güzeller, C. O. (2011). A study of cross-cultural equivalence of computer attitude in PISA 2009 student questionnaire. *Education and Science, 36*(162), 320-327.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage.

IEA (2016). *TIMSS 2015 Assessment frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center* website: https://timss.bc.edu/timss2015/frameworks.html Date of access: 5 January 2018.

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

Kıbrıslıoğlu, N. (2015). *The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey - China (Shangai) - Indonesia* (Master Thesis). https://tez.yok.gov.tr accessed from.

Köse, İ. A. (2015). Investigation of items in PISA 2009 student questionnaire subscales (q32-q33) in terms of differential item functioning. *K*astamonu Education Journal, 23(1). 227-240

Magis, D., Beland, S., & Raiche, G. (2016). *DifR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics.* R package version, 4.7

Mantel N., & Haenszel W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease*. Journal of The National Cancer Institute, 22*(4), 719-748. https://doi.org/10.1093/jnci/22.4.719

MEB (2016). *TIMMS 2015 Ulusal matematik ve fen bilimleri ön raporu 4. ve 8. Sınıflar.* MEB, Ölçme ve Değerlendirme Sınav Hizmetleri Genel Müdürlüğü (Online) http://timss.meb.gov.tr/wp-content/uploads/TIMSS_2015_Ulusal_Rapor.pdf Date of access: 5 March 2018.

Murtin, M.O, Mullis, I.V.S., Foy, P. &, Hooper, M. (2015). *TIMSS 2015 International Results in Science*. http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Science-Grade-8.pdf Date of access: 15 February 2018.

Osterlind, S. J. (1983). *Test item bias* (Ed. 30). Sage.

Osterlind, S. J., and Everson, H. T. (2009). *Differential item functioning* (2.ed.). Sage

Peterson, R.A. (1979). Revitalizing the culture concept. *Journal of Annual Review of Sociology, 5*, 137-166.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Journal of Applied Psychological Measurement, 14*(2), 197–207. https://doi.org/10.1177/014662169001400208

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Team, R. (2013). R development core team. *RA Lang Environ Stat Comput, 55,* 275-286.

Toprak, E. ve Yakar, L. (2017). Analysis of SBS 2011 Turkish subtest items in terms of differentıal item functioning by different methods, *International Journal Of Eurasia Social Sciences, 8*(26), 220-231

Uyar, Ş. & Kaya Uyanık, G. (2016). Investigation of PISA 2012 cognitive items in terms of differential item functioning based on culture. *Journal of Research in Education and Teaching, 5*(3), 230-240.

Uzun Başusta, N. B. (2010). Measurement invariance. *Journal of Measurement and Evaluation in Education and Psychology, 1*(2), 58-64*.*

Uzun, N.B., & Gelbal, S. (2017). An investigation of item bias in PISA science test in terms of the language and culture. *Kastamonu Education Journal, 25*(6), 2427-2446.

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. https://doi.org/10.1080/15434300701375832

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* National Defense Headquarters.