Cypriot Journal of Educational Sciences

Volume 13, Issue 2, (2018) 375-384

www.cjes.eu

The comparison of differential item functioning predicted through experts and statistical techniques

Nuri Dogan[*]**,** Hacettepe University, Faculty of Education, Beytepe Mahallesi, Çankaya, Ankara 06800, Turkey.
Ronald K. Hambleton, University of Massachusetts, Amherst, MA 01003, USA.
Meltem Yurtcu, Hacettepe University, Faculty of Education, Beytepe Mahallesi, Çankaya, Ankara 06800, Turkey.
Sinan Yavuz, University of Wisconsin - Madison, Educational Psychology Department, Madison, WI 53706, USA.

Suggested Citation:
Dogan, N., Hambleton, R. K., Yurtcu M. & Yavuz, S., (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Science*. *13*(2), 375–384.

Received date October 19, 2017; revised date February 07, 2018; accepted date June 05, 2018.
Selection and peer review under responsibility of Prof Dr. Huseyin Uzunboylu Near East University.
©2018 Academic World Education & Research Center. All rights reserved.

Abstract

Validity is one of the psychometric properties of the achievement tests. To determine the validity, one of the examinations is item bias studies, which are based on Differential Item Functioning (DIF) analyses and field experts' opinion. In this study, field experts were asked to estimate the DIF levels of the items to compare the estimations obtained from different statistical techniques. Firstly, the experts were asked to examine the questions and make the DIF level estimations according to the gender variable for the DIF estimation, and the agreement of the experts was examined. Secondly, DIF levels were calculated by using logistic regression and the Mantel-Haenszel (MH) statistical method. Thirdly, the experts' estimations and the statistical analysis results were compared. As a conclusion, it was observed that the experts and the statistical techniques were in agreement among themselves, and they were partially different from each other for the Sciences test and equal for the Social Sciences test.

Keywords: Item bias, differential item functioning (DIF), expert estimation.

Nuri Dogan*, Hacettepe University, Faculty of Education, Beytepe Mahallesi, Çankaya, Ankara 06800, Turkey.
E-mail address: nurid@hacettepe.edu.tr / ***Tel.:*** +90-312-297-8550

1. Introduction

Impartiality is a sing for validity (Camilli & Shepard, 1994). So, one of the validation studies, the application of which has become a routine in recent years, is the item bias studies. Item bias studies mostly cover a review of sensitivity and Differential Item Functioning (DIF) (ETS, 2009; Hambleton, 2006; Sireci & Mullane, 1994). The sensitivity review and DIF studies contribute to the achievement test scores in a proper manner (Zieky, 2002). DIF, for the achievement tests, is defined as the differentiation in the probabilities of giving the correct answer to an item on the part of the individuals with the same competency level who belongs to different groups of the same population (Hambleton & Rogers, 1996; Zumbo, 1999). The fact that the probability of giving the correct answer to the item for different ability level students is expected to be different. Moreover, the probability of giving the correct answer to the item is supposed to be equal for the individuals with the same ability level who exist within the same population, even though they belong to different groups. If the levels of the students, who are within the same population but belong to separate groups, have a different probability of giving the correct answer to the item, it is considered to be the item bias (Zumbo, 1999). Items with bias indication interfuse different variables to the assessment process other than examiners ability (Cromwell, 2002).

The total test scores, which contain biased items, will be non-objective, and the decision made according to the overall scores will be faulty and misjudged. As a result, the validity of the assessment will fail. At the beginning of the statistical analyses, the important point is to determine which variable the DIF analyses will be performed on (Hambleton & Rodgers, 1996; Hambleton, Ying & Klauck, 2001). DIF analyses can be carried out by taking different variables into consideration, such as gender, ethnic group, social class, subculture, beliefs, etc. Another important respect regarding these analyses is the stage of determining the test hypothesis and statistical techniques. DIF analyses can be performed by applying statistical techniques based on the classical test theory (CTT) and the item response theory (IRT). Even if the CTT hypothesis is sample dependent, CTT techniques are still more practical than IRT methods (Budgell, Raju & Quartetti, 1995; Hambleton, 2006; Hambleton & Rogers, 1989; Jones & Hambleton, 1992). In this research, two statistical techniques, based on the CTT, were applied.

The decision on the superiority or importance of DIF values is made according to the results obtained from the statistical analysis. Since the statistical significance is influenced by the sample size of the test, calculation of the effect size is also becoming more and more common (Benito, Hidalgo & Guilera, 2010; Hambleton, 2006). While DIF is an item analysis methodology that describes the sample as a whole, ignoring how the psychometric properties of the scale may vary as a function of variation within the sample (Zumbo,1999) because it uses a variety of techniques. The techniques, such as the Mantel-Haenszel (MH) statistical method and logistic regression, indicate the effect size quantity for the calculated DIF value (Hambleton, 2006). Hence, the items in the MH method and logistic regression techniques can be grouped as those yielding weak (A), moderate (B) and high (C) DIF levels by making use of the effect size quantity. The final step of the analytical processes is about determining, in favour or in disfavour of, which group the DIF values prove to be. At this stage, the DIF type of the items is identified as uniform or non-uniform (Mellenbergh, 1982). In the uniform DIF-yielding items, the item functions are favourable for either the reference or the focal group on all ability levels. Whereas the non-uniform DIF-yielding item is advantageous for the reference group on some ability levels and it is favourable for the focal group on different ability levels.

It is rather challenging to make an interpretation about item bias and fairness of the test through statistical techniques (Zieky, 2002). An item might have yielded DIF for some reason other than item bias (Camilli & Shepard, 1994). Apart from the fact that a study on DIF is considered to be one of the evidence-gathering ways for the validity of the test, since such evidence have no single correct answer, an expert or a referee advice is required to evaluate and interpret this evidence (Benito et al., 2010).

Hambleton and Rogers (1995) emphasised that the sensitivity reviews could be beneficial if it is done before performing a statistical analysis to determine whether or not the items have a structure that is in favour or disfavour of an aggressive, controversial and particular group. The aim of the sensitivity review is to reveal the source of DIF in the items after statistical analyses.

To determine the items, which contain the expressions likely to cause bias, the experts try to examine whether these items bear stereotyped expressions or not. Whether the content is unfavourable for the experiences of a given group or not, the sub-groups have equal chances regarding learning the substance of the item(Benito et al., 2010; Hambleton et al., 2001). In this evaluation, education and training of DIF and item bias can be provided to avoid any difference among the experts and elevate their adaptability (Hambleton, 2006). It is hard to predict bias, even though differences among the experts in regard to their training might have been made up (Gierl, Rogers & Klinger, 1999; Jensen, 1977; Plake, 1980; Sandoval & Miille, 1980).

There are rather few studies found in the literature to determine the items with DIF and to understand the sources of DIF by depending on the experts' view (Gierl et al., 1999; Roth, Oliveri, Sandilands, Lyons-Thomas & Ercikan, 2013). The studies regarding the experts' predictions without having any knowledge of statistic results and the results of the statistical DIF-determining techniques are rather insufficient.

In the same way, no study has been found in the literature as to the comparison of the predictions of the field experts, who lack sufficient level of theoretical knowledge even if they may have received training on DIF.

This research was designed for the purpose of eliminating such imperfections. The primary goal of the study is to compare the results of the experts' predictions on DIF and the statistical results. In line with this purpose, the explanations to the following questions were sought for:

1. How are the DIF level predictions of the field experts for the Sciences and Social Sciences tests?
2. How are the DIF predictions made through statistical techniques for the Sciences and Social Sciences tests?
3. How are the comparison results between the DIF level-predictions addressed by the field experts and statistical methods results for the Science and Social Sciences tests?

2. Method

2.1. Sample and data

This research has been designed as a descriptive study since it is aimed at putting forward the compliance between the DIF predictions of the field experts and the DIF results calculated through the statistical techniques. The research population comprises 1,055,508 eighth grade students who entered the Placement Test (PT, the Turkish acronym is SBS) performed by the Ministry of National Education of Turkey (MNE). A total of 130,564 students were selected from this population through the unbiased methods, and they were incorporated into the sampling within the scope of the research. During the sampling process, 13% of the population were selected randomly (draw technique); yet, those who left their gender unwritten and those whose PT (SBS) scores proved to be zero were excluded from the selected sampling. As a result, the analyses were conducted with 130,564 students: 65,505 of whom were male, and 65,058 of whom were female. In the first stage of the research, the items and test scores within the PT Sciences and Social Sciences tests performed on the eighth-grade students by MNE in 2011 were practiced. These data were provided from the MNE. The PT carried out for the eighth-grade students consists of five tests: Turkish, Mathematics, Sciences, Social Sciences and Foreign Languages. The entire test performed by the eighth-grade students is composed of a total of 100 questions, comprising 23 questions in the Turkish test, 20 questions in the Mathematics test, 20 questions in the Sciences test, 20 questions in the Social Sciences test and 17

questions in the Foreign Languages test. The exam duration is 120 minutes. The study was implemented according to the data obtained from the Sciences and Social Sciences tests.

2.2. DIF-determining techniques (DIF procedures)

To determine DIF levels, PT Sciences and Social Sciences tests items and the statistical techniques regarding the decisions made by an expert (e.g., the judgemental technique) were used. To ascertain the DIF levels of the items with a statistical approach, the techniques referred to as the 'Mantel-Haenszel (MH)' method and the 'indices of conditional p-value differences' were applied. MH is a Chi-Square statistic, and the obtained results can be interpreted as DIF in favour of the reference group if MH > 1; DIF in favour of the focal group if MH < 1; and no DIF if MH ≅ 1. A logarithmic transformation is performed to be able to interpret the MH statistics more easily. The logarithmic transformation formula is as follows: $\Delta\Omega = \Delta MH = -(4/1.7)*\ln MH = -2.35*\text{logit}$. The results obtained by the logarithmic conversion of the formula are interpreted as DIF in favour of the focal group if $\Delta MH > 0$; DIF in favour of the reference group if $\Delta MH < 0$; and no DIF if $\Delta MH \cong 0$ (Holland & Thayer, 1986).

Separately, the DIF level can also be interpreted according to the size or greatness of MH. It is stated that if $|\Delta MH| < 1$, then A indicates an insignificant DIF level. If $1 \leq |\Delta MH| < 1.5$, then B indicates a moderate DIF level. Thus, if $|\Delta MH| \geq 1.5$, then C indicates a high DIF level (Dorans & Holland, 1993). One of the weakest aspects of this technique is that it cannot distinguish the uniform DIF and non-uniform DIF from one another. The second statistical technique, referred to as the 'indices of conditional p-value differences' and used in defining the item bias within the Sciences and Social Sciences tests, is also termed as the standardised differences, or merely, standardisation differences. In this technique, the test takers within the reference and focal groups were equalised according to the total test scores and score categories in the first place. Afterward, the difference among the percentages of correct items of the focal and reference groups was taken and standardised for each equalised score. The positive values obtained from the analyses suggest that the DIF is in favour on the reference group, whereas the negative values indicate that the DIF is in favour on the focal group. If the DIF statistics obtained is within the range of ±0.5, then DIF is regarded as insignificant. If the DIF statistics falls out of that range, then DIF is considered to be significant. Signed and Unsigned DIF (SDIF and UDIF) statistics can be calculated through the standardisation method. If the difference between the two statistics is small, then the existence of a uniform DIF is considered; whereas, if it is large, then a non-uniform DIF is mentioned. On the other hand, in the process of consulting the experts' decisions in determining DIF, DIF predictions of the field experts, which were made according to gender for the Sciences and Social Sciences tests, were also collected. The volunteerism of all the experts for their participation in this research was taken as the cornerstone. The field experts consisted of the teachers who had completed the involved undergraduate program as well as those experienced in at least a two-year teaching process. Seven field experts of sciences (i.e., three males and four females) and five field experts of social sciences (i.e., two males and three females) have performed their DIF predictions.

First of all, during face-to-face interviews, the experts of sciences and social sciences were informed as to what DIF was and how it could be identified. In addition, the experts were shown sample questions with weak, moderate and high DIF levels determined in different studies. The field experts were given tests regarding their area of expertise, and then they were asked to fill in the form given to them by marking the DIF level of the items within the test as either weak (A), moderate (B) or high (C).

The field experts were asked to use a list of nine items adapted from Hambleton and Rogers (1996) and Hambleton et al. (2001), in regard to providing assistance for the DIF predictions to be performed according to gender. These items are as follows:

- 'There is some content likely to arouse different emotions or cause fallacy according to gender'.
- 'It contains structure and language bearing vulgar and insulting characteristics according to gender'.
- 'There is some content showing difference according to gender'.

- 'It contains a structure/structures that the individuals may take advantage of in their lives according to their sexual identities'.
- 'It contains the information male/female students can benefit from'.
- 'It contains words, structures or situations causing differences in meaning for female/male students'.
- 'For cultural reasons, the distractor(s) differ(s) according to gender'.
- 'The explanations given within the question may cause confusion in students' minds according to gender'.
- 'It contains a hint/clue in the way that it will be of use to female/male students'.

The predictions of the field experts for the DIF level of each question (i.e., weak (A) or no level, moderate (B) level or high (C) level) were tabulated within an Excel file. These predictions included in the classification scale were evaluated following the mode value. Attempts to solve the research problems were conducted by comparing the DIF predictions obtained from the field experts through statistical techniques.

3. Findings

In this study, 65,505 male and 65,088 female students answered the tests. When the descriptive statistics of the Sciences test were examined, the females had a mean test score of 9.04, and the males had a mean test score of 8.18. For the Social Sciences test, the females had a mean test score of 9.74, and the males had a mean test score of 8.68. It can be noted that female students are more successful than male students for both tests. When these means are compared to the Sciences test, the difference between female and male mean test scores is significantly important on behalf of the female students (t = 30.827, p < 0.001). The same results were also found for the Social Sciences test (t = 31.408, p < 0.001). The estimations made by the field experts by using a list of DIF indicators to evaluate the items in the Sciences and Social Sciences tests were collected as a reply to the first question of the study.

Table 1. The DIF estimations of the field experts on sciences test items

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Field Experts	A1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	1	2	1	1	1	1
	A2	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1
	A3	1	1	1	2	2	1	2	1	1	1	2	1	1	1	1	2	2	1	1	1
	A4	1	1	2	2	2	1	2	1	2	1	2	1	1	1	2	2	1	1	1	2
	A5	1	1	2	2	2	1	2	1	1	1	2	1	1	1	2	2	1	1	2	1
	A6	1	1	2	2	1	1	1	1	1	1	1	1	1	1	2	2	2	1	2	2
	A7	1	1	2	2	2	1	1	1	1	2	2	1	1	1	2	2	1	1	2	2
1	%	100	100	43	14	43	100	43	100	86	57	43	100	100	100	43	0	57	100	57	57
2	%	0	0	57	86	57	0	57	0	14	43	57	0	0	0	57	100	43	0	43	43
3	%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Field experts' answers; 1: Weak level (A), 2: Moderate level (B) and 3: High level (C).

As observed in Table 1, DIF level estimations of the field experts are highly close on many items for the Sciences test. Percentage values show that a number of items present DIF at a weak level much more than the other DIF levels. For the 16th item, all of the field experts stated that it showed DIF at a medium level. The field experts did not find any of the items as being at a higher-level DIF. The field experts expressed that the finding of DIF in the Sciences test items was due to the following reasons: 'It contains information that may be benefited by male/female students'; 'The distractor/s show differences in terms of cultural reasons according to gender'; 'It contains structure/s that may be advantageous in their lives according to the gender identities of the individuals'. The agreement of the replies given by the seven field experts was checked with the percentage agreement of Krippendorff.

The measured value was found as 0.714. The DIF estimations of the field experts on the items of the Social Sciences test are given in Table 2.

Table 2. The DIF estimations of the experts of the field on social sciences test items

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Field Experts	S1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1
	S2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	S3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	S4	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
	S5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	%	100	100	100	100	100	100	100	80	100	80	100	100	100	100	100	100	100	100	100	100
2	%	0	0	0	0	0	0	0	20	0	20	0	0	0	0	0	0	0	0	0	0
3	%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Field experts' answers; 1: Weak level (A), 2: Moderate level (B) and 3: High level (C).

As observed in Table 2, DIF level estimations of the field experts, on the Social Sciences test, show that two of all items have a weak DIF level. Only two experts expressed medium-level DIF for two different questions. According to the estimations made by experts by using a list with DIF indicators, the experts stated that these two items were considered under the following item: 'The distractor(s) show(s) differences regarding cultural reasons according to gender'. The agreement of the replies given by the five experts was checked with the percentage agreement of Krippendorff. The measured value was found as 0.96.

The DIF levels were calculated for the Sciences and Social Sciences test items by using the MH and standardisation technique for the second question of the study. SDIF and UDIF indices are calculated by the standardisation technique. SDIF explains the weighted average difference between the reference and focus group and estimates single DIF value for items (Dorans & Kulick, 1986). If the focus group has a non-uniform and uniform DIF, UDIF measures the difference between item *p*-values of the reference group. In these regards, the effect of the sample size difference between groups are reduced to a minimum. On the other hand, item effect can be statistically tested by the MH method. The goal by these two methods is to obtain the distribution of DIF items. The DIF results calculated according to the mentioned methods and results are given in Table 3.

According to the results of the MH Chi-square statistics for the Sciences test in Table 3, all the other items were found to be significant at an 0.05 level, except for items 6, 11, 12 and 19. However, when the MH D-DIF (Mantel-Haenszel differential item functioning) results were examined and corrected by considering the size of the sampling, it was observed that none of the items gave DIF at a significant level. According to the MH method, all items have DIF at a weak level.

The SDIF and UDIF values that were calculated with the standardisation technique are highly close to each other, and these values vary between -0.041 and 0.046. When the SDIF and UDIF values that were computed with the standardisation technique are examined, we find out that none of the items have DIF.

Table 3. The MH and standardisation DIF statistics calculated for gender for sciences and social sciences tests

	Sciences					Social sciences				
		MH		Standardisation			MH		Standardisation	
Item	Alpha	DIFL	MH D-DIF	SDIF	UDIF	Alpha	DIFL	MH D-DIF	SDIF	UDIF
1	1.154	A	−0.337	0.027	0.028	0.923	A	0.188	−0.014	−0.022
2	1.16	A	−0.349	0.027	0.028	0.971	A	0.07	−0.003	−0.01
3	0.919	A	0.199	−0.012	−0.015	1.027	A	−0.064	0.003	0.023
4	1.064	A	−0.146	0.01	0.012	1.022	A	−0.051	0.005	0.013
5	0.913	A	0.213	−0.016	−0.018	1.015	A	−0.035	0.002	0.016
6	1.001	A	−0.002	0.001	0.007	1.31	A	−0.635	0.04	0.04

7	1.07	A	−0.158	0.011	0.015	1.292	A	−0.602	0.038	0.038
8	1.306	A	−0.628	0.046	0.046	1.02	A	−0.047	0.003	0.011
9	0.939	A	0.147	−0.011	−0.021	0.904	A	0.238	−0.011	−0.02
10	0.74	A	0.707	−0.036	−0.041	0.947	A	0.128	−0.006	−0.022
11	0.976	A	0.056	−0.004	−0.012	0.907	A	0.228	−0.012	−0.015
12	0.979	A	0.05	−0.003	−0.012	0.995	A	0.011	−0.001	−0.011
13	1.081	A	−0.184	0.014	0.017	0.883	A	0.291	−0.022	−0.023
14	1.127	A	−0.28	0.02	0.024	1.107	A	−0.24	0.017	0.019
15	0.866	A	0.338	−0.02	−0.023	1.127	A	−0.28	0.017	0.017
16	0.814	A	0.483	−0.04	−0.04	0.963	A	0.089	−0.006	−0.019
17	0.949	A	0.124	−0.01	−0.017	1.055	A	−0.125	0.007	0.018
18	0.941	A	0.142	−0.011	−0.023	0.832	A	0.432	−0.018	−0.024
19	0.997	A	0.008	0	0.007	0.851	A	0.379	−0.017	−0.026
20	1.045	A	−0.102	0.008	0.013	0.854	A	0.372	−0.024	−0.028

DIFL = DIF level.

According to the results of the MH Chi-square statistics computed for the Social Sciences test in Table 3, all the other items were found to be significant at an 0.05 level, except for items 3, 4, 5, 8 and 12. When the MH D-DIF results were examined and adjusted by considering the size of the sampling, it was seen that none of the items gave DIF at a significant level. According to the MH method, all items have a weak (A) DIF level.

The SDIF and UDIF values that were calculated with the standardisation technique are highly close to each other, and these values vary between -0.028 and 0.040. When the SDIF and UDIF values were tested and measured with the standardisation technique, we found that none of the items have an important DIF level.

DIF level estimations were made for the Sciences and Social Sciences test items to answer the third question of the study, and the DIF level estimations obtained from the Statistical Techniques were compared.

All the items that were calculated for the Sciences test and estimated according to the MH method showed DIF at a weak level. Meanwhile, according to the UDIF values that were calculated with the standardisation technique for the Sciences test, we found that none of the items have DIF. When the experts were asked to make estimations according to a list in which the DIF indicators were given, it was remarked that the majority of the DIF levels of the items in the Sciences test were a weak (A) level. According to all of the experts, the 16th item in the Sciences test showed a moderate (B) DIF level . According to 86% of the experts, the 4th item presented a moderate (B) DIF level. Likewise, 57% of the experts found that the 3rd, 5th, 7th, 11th and 15th items presented a moderate (B) DIF level. According to these findings, the number and level of the items with DIF are different, which was discovered by statistical techniques and based on the estimations of the experts. The reason for these differences could be the sensitivity of the check lists, which were giving to the field experts. Items 6 and 12 were interpreted by the experts as non-DIF items, even though these items have a statistically weak (A) DIF level. It is also possible that this problem may occur because of some other variable than gender (e.g., students might not understand these items correctly). For the item 11, which is determined as significant but has a weak (A) DIF level by statistical techniques, field experts found a moderate (B) DIF level. In this case, we can say that distracters of the item worked similarly for both female and male students. On the other hand, other than item 19 and the aforementioned items, field experts' opinions and statistical results coincide.

All of the items that were calculated for the Social Sciences test and estimated according to the MH statistical method showed a weak (A) DIF level. Meanwhile, according to the UDIF values calculated with the standardisation technique for the Social Sciences test, it is found that the investigated items

do not show DIF. When the experts were asked to make estimations according to a list in which the DIF indicators were given, it was observed that the majority of the DIF levels in the Social Sciences test were at a weak (A) level). It is apparent that only two items in this test showed DIF at A-level according to the 80% of the experts, however all of the items showed DIF at A-level by statistical analyses. According to these results, the number and level of the items with DIF, which was determined with statistical techniques, and that were based on the estimations of the experts of the field are the same. For items 3, 4, 5 and 12, the results were statistically significant ($p < 0.05$), but the items did not have an important DIF; thus, there could be some other variables that effect students' answer patterns. For item 10, field experts mentioned that gender might have a slight effect. Aside from these items, however, all the statistical analyses and opinions of the field experts overlap.

On the other hand, it is possible to suggest that the agreement between the field experts in the Social Sciences test and the agreement between the experts and the statistical technique results is higher than that of the Sciences test.

4. Results

The estimations made by the experts were studied by using a list having DIF indicators in nine items to evaluate the Sciences and Social Sciences tests. It is noted that the majority of the items have a weak (A) DIF level, and the rest of the items have a moderate (B) DIF level. Meanwhile, according to the experts, none of these items have a high (C) DIF level. The field experts stated that the finding of DIF levels in seven Sciences test items as moderate (B) was due to the following reasons:

- 'It contains information that may be benefited by male/female students'.
- 'The distractor/s show differences in terms of cultural reasons according to gender'.
- 'It contains structure/s that may be advantageous in their lives according to the gender identities of the individuals'.

Statistical values, such as the Krippendorff Alpha and Fleiss Kappa, were not obtained, but the percentage agreement was calculated instead because the estimations of the experts were not very different. The experts did not prefer the expression, 'It contains rude or insulting structures or language', as a DIF indicator in the list. The agreement of the replies given by the seven experts to the Sciences test items was reviewed with the percentage agreement. The measured value was calculated as 0.7142.

According to the experts, all of the Social Sciences test items have a weak (A) DIF level. According to 20% of the experts, who made estimations by using a list of DIF indicators, affirmed that only two items represented the following reason: 'The distractor/s show differences in terms of cultural reasons according to gender'. Social Sciences, which is a verbal subject, may be defined as a discipline that increases the readiness of the individuals in society. In this situation, it can be claimed that there were no items according to gender with DIF in PT Social Sciences test in 2011. The agreement of the responses given by the five experts in Social Sciences was checked with the percentage agreement. The value was found as 0.96. This value shows that there is a remarkably high agreement between experts.

The DIF levels were calculated for the Sciences and Social Sciences test items by using the MH and standardisation technique. The D-DIF results, according to the MH technique for Sciences and Social Sciences, show that the items in these tests have a weak (A) DIF level. In a similar large-scale examination conducted in Turkey, there are items with DIF in Sciences test, and this has been demonstrated in the literature (Kalaycıoğlu & Kelecioğlu, 2011; Yurdugül & Aşkar, 2004). Similarly, when the calculations were made for both tests by using the standardisation technique, the SDIF and UDIF values indicate that all the items show a weak (A) DIF level.

The DIF level estimations made by the experts for the items in Sciences and Social Sciences test and the DIF level estimations obtained from the statistical techniques were compared. All the items estimated according to the MH statistical method and standardisation technique for the Sciences test showed a weak (A) DIF level. When the experts were asked to make estimations according to a list, in which the DIF indicators were given, it was observed that more than 50% of the field experts stated that items 3, 4, 5, 7, 11, 15 and 16 in the Sciences test showed a moderate (B) DIF level. It was asserted that the DIF level of the other items was weak. According to these results, although the number of the items with DIF, both regarding statistical techniques and experts' opinions are partly different. All the items estimated according to the MH statistical method and standardisation technique for the Social Sciences test showed a weak (A) DIF level.

When the experts were asked to make estimations according to the list in which the DIF indicators were given, it was observed that all of the items in the Social Sciences test showed a weak (A) DIF level. According to these results, the number of items with DIF determined with statistical techniques and the estimations of the experts is equal. It is possible to suggest that there is a full agreement between them. It can be indicated that the agreement between the experts for the Social Sciences is higher than the agreement between the results of the experts' decisions and statistical techniques in the Sciences test. This situation might get effected by the different numbers of field experts. On the Sciences test, having a large number of experts would cause a low level of consistency between the experts. In other words, it would increase the heterogeneity. However, it is mentioned in literature that in the 2011 PT Sciences test, DIF items were found, but there were none found in the Social Sciences test (Kan & Sünbül ve Ömür, 2013). Although, in the process of selecting the experts, we tried to select an equal number of male and female experts to decrease the chance of bias.

We can further express that the main reason of the difference between the field experts and the statistical techniques was allowing the experts to control the very detailed check list. For future studies; instead of having the difference in the subgroups by gender, focusing on other possible DIF causing variable may give different results. We can also draw a conclusion such as; expert opinion is crucial in the field of social sciences, which centres human being, to determine bias on measurement applications, and specifying DIF items. However, even if the items in the check list are selected meticulously, it can give more detailed information than statistical techniques. With this study, the number of studies will increase in the literature.

The techniques, which are statistically based on the CTT and uses structurally similar tables, give very similar results. This finding overlaps with Ward and Bennett's (2012) study. In addition, for a $p < 0.05$ significance level, the MH method was found to be more sensitive than the standardisation technique because the MH method is sensitive to the bias of distractors (Kurnaz, 2005). However, according to Selvi (2013), the standardisation technique determines more DIF items than the MH method.

Acknowledgements

This study has been realised in the process of 'Comparison of Classical Test Theory and Item Response Theory regarding Test Development Process', which is carried on by Nuri Dogan and is supported in the scope of TUBITAK BIDEB 2219 'Post-doctorate Research Scholarship Program Abroad' with the number 1059B191400868.

References

Bakan Kalaycıoğlu, D. ve Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. Eğitim ve Bilim, 36 (161), 3-12

Benito, J. G., Hidalgo, M. D. & Guilera, G. (2010). Bias in measurement instruments: Fair tests. Papeles del Psicólogo. 3, 1, 75-84.

Budgell, G. R., Raju, N. S. & Quartetti, D. A. (1995). Analysis of Differential item functioning in translated assessment instruments. Applied Psychological Measurement. 1, 4, 309-321.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Volume 4, Thousand Oaks: Sage Publications.

Cromwell, S. (2002). Primer on ways to explore item bias. Paper presented at the 25th Annual Meeting of the Southwest Educational Research Association14-16 February, Austin, Texas, USA.

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. Educational Testing Service. Retrieved May 22, 2015, from http://www.ets.org/Media/Research/pdf/RR-13-27.pdf

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel Haenszel and standardization. In P. W. Holland, ve H. Wainer, (Eds.), Differential Item Functioning (pp. 35–66), New Jersey: USA.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Educational Testing Service. Guidelines for fairness review of assessment. Retrieved May 22, 2015, from http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Gierl, M.J., Rogers, W.T. & Klinger, D.A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. The Alberta Journal of Educational Rese. 45, 4, 353-376.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. Medical Care. 44, 11, 182-188.

Hambleton, R. K. & Jones, R. W. (1992). Comparison of empirical and judgmental methods for detecting differential item functioning. Paper presented at the Annual Meeting of the National Council on Measurement in Education. April 21-23, SanFrancisco, CA, USA.

Hambleton, R. K. & Rogers, H. J. (1995). Item bias review. Practical Assessment, Research and Evaluation. 4,6. Retrieved July 9, 2015, from http://pareonline.net/getvn.asp?v=4&n=6.

Hambleton, R. K. & Rogers, H. J. (1996). Developing an item bias review form. Retrieved May 22, 2015, from http://ericae.net/ft/tamu/biaspub2.htm

Hambleton, R.K. & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education. 2, 4, 313-33.

Hambleton, R. K., Ying, L. & Klauck, S. (2001). Pennsylvania department of education item sensitivity review procedure. Unpublished document.

Holland, P.W., & Thayer, D.T. (1986). Differential item performance and the Mantel-Haenszel procedure. (Technical Report No. 86–69). Princeton, NJ: Educational Testing Service.

Jensen, A. R. (1977). An examination of cultural bias in the Wonderlic Personnel Test. Intelligence, 1, 51-64.

Jones, R. W. & Hambleton, R. K. (1992). Recent advances in psychometric methods. Laboratory of Psychometric and Evaluative Research Report No. 233.Amherst, MA: University of Massachusetts, School of Education.

Kan, A., Sünbül, Ö & Ömür, S. 6. - 8. Sınıf Seviye Belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi. Mersin Üniversitesi Eğitim Fakültesi Dergisi, Cilt 9, Sayı 2, Ağustos 2013, ss.207-222.

Kurnaz, F.B. (2006). Peabody Resim Kelime Testinin Madde Yanlılığı açısından incelenmesi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Eğitim Bilimleri Anabilim Dalı. Yüksek Lisans Tezi, Ankara.

Mellenbergh, G. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics. 32, 1, 92-109.

Plake, B. S. (1980). A comparison of statistical and subjective procedures to ascertain item validity: one step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., and Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. International Journal of Science Education, 35, 546–576.

Sandoval, J., & Miille, W. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

Selvi, H. (2013). Klasik test ve madde tepki kuramlarına dayalı değişen madde fonksiyonu belirleme tekniklerinin farklı puanlama durumlarında incelenmesi. Mersin Üniversitesi, Eğitim Bilimleri Enstitüsü, Yayınlanmamış Doktora Tezi, Mersin.

Sireci, S. G. & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: The sensitivity review process. CLEAR Exam Review. 5, 2, 22-27.

Ward, W.C., Bennett, R.E. (2012). Construction versus choice in cognitive measurement: ıssues in constructed response, performance testing, and portfolio assessment. Routledge,Taylor ve Francis Group, London and New York.

Yurdugül, H. & Aşkar P. (2004). Ortaöğretim Kurumları Ögrenci Seçme ve Yerleştirme Sınavı'nın cinsiyete göre madde yanlılığı açısından incelenmesi (The investigation of the student selection and placement examination for secondary education with respecet to gender in terms of item bias). Egitim Bilimleri ve Uygulama Dergisi, 3(5), 3-20.

Zieky, M. (2002). Ensuring the fairness of Licensing Tests. CLEAR Exam Review. 12, 1, 20-26

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. Canada: Ottowa, Directorate of Human Resources Research and Evaluation National Defense Headquarters: Author.