World Applied Sciences Journal 3 (2): 231-239, 2008
ISSN 1818-4952

Investigation of Peabody Picture Vocabulary Test from the Point of Item Bias Peabody Picture Vocabulary Test

[1]*Fatma Betül Kurnaz and* [2]*Hülya Kelecioğlu*

[1]Educational Sciences Faculty, Ankara University, Ankara, Turkey
[2]Education Faculty, Hacettepe University, Ankara, Turkey

Abstract: In this research, differential item functioning (DIF) of the items included in Peabody Picture Vocabulary Test (PPVT) were analysed according to socio-economic level and gender. Responses of 592 children to PPVT were used within the settings of this research. Mantel-Haenszel (MH) and logistic regression (LR) methods were utilized and the items displaying DIF were analysed with distractor response analyses. It was found out that there are items displaying DIF according to both methods in gender and socio-economic level but a harmony could not be obtained between these two methods. It was observed that the items displaying DIF decreased the internal consistency (KR-20). The findings of the research revealed that MH method was much more sensitive to distractor bias and was also able to determine the bias originating from distractors.

Key words: Differential Item Functioning · Bias · Mantel-Haenszel · Logistic Regression

INTRODUCTION

Bias is defined as systematic error within measurement process and causes score distribution of subgroups that were given test differ from each other [1,2]. In the relevant literature, two types of bias were mentioned [1-3] these are internal and external bias:

Internal bias results from the psychometric characteristics of the test items and it is dependent on the test and its content. The bias that does not result from test and test content and the bias which is related to the conditions of testing is called as *external bias*. Individual and group differences can also cause a systematic bias on test scores and this bias is defined as external bias. If there is a difference between groups from the point of variable measured, this difference belongs to groups. In this case, to reach the conclusion that the test or the item is biased can be wrong.

Besides, matters that can cause biases are also dwelt on in the literature. Ackerman [4] stated that the difference in the performance of a measured variable of two groups, unequalness of standard deviation in two groups and unequalness of the correlation between "the valid and nuisance dimensions" for two groups can cause a bias. Moreover, Van de Vijver and Portinga [5] mentioned that many issues such as the structure functions differently between groups, inappropriateness of the sample for comparison, social acceptance, item type, having a difference in application conditions and lack of strength resulting from translation can cause biases.

Differential item functioning (DIF) has a meaning different from the item bias. If the probability of answering the item correctly differs for two groups having same ability level, it can be said that there is a DIF in the item. Obtaining many DIF in the items threaten the validity of the test scores and can cause wrong results in interpreting the scores of two groups. However, obtaining an item displaying DIF does not definitely indicate that the item is biased. It is possible to determine the bias of an item with DIF, but on the other hand, the ongoing processes should also be observed throughout the duration of decision making whether an item is biased. These processes are content analysis, empirical evaluation and specialist opinion [6-8].

Allalouf [9] stated that DIF analysis would guide people who adapted the test and increase the validity of the test in adapting a test from one language to another and also mentioned that a successful revision has four objectives. These are; (a) replacing the items translated instead of removing from the test, (b) finding out the origin of DIF, (c) determining the DIF within the items written in different type, (d) increasing validity by decreasing DIF within the items revised.

Corresponding Author: Dr. Fatma Betül Kurnaz, Educational Sciences Faculty, Ankara University, Ankara, Turkey

There are two types of DIF, uniform DIF and nonuniform DIF. When the difference in the probability of answering the item as correct by two groups having the same skill level becomes constant, DIF is uniform. If the difference in the probability of answering the item as correct by two groups having the same skill level does not become constant from the points of amount and direction, there is a nonuniform DIF in the item.

The methods used in detecting the item bias are classified under three categories Allalouf [9]:

1. Item response theory methods
2. χ^2 methods
3. Item difficulty methods

In item response theory (IRT) methods, when the item characteristic curves (ICC) of all subgroups become same, it is decided that the item is unbiased. When the item characteristic curves become same, it is considered that the items measure the same latent feature.

In χ^2 methods, when the probability of answering the item as correct by all people having the same ability level without depending on any cases such as age, gender or race becomes equal, the item is accepted as unbiased. However, if the probability of answering the relevant item as correct that belongs to respondents in an item display differences, it is suggested to carry out a distractor response analysis which is one of the way of determining bias. If the answers given by respondents to the distractors are significantly differs in an item, it is concluded that there is a bias in an item. Logistic regression (LR), Mantel-Haenszel (MH) and Distractor Response Analysis methods considered in this research are also χ^2 methods [2].

In item difficulty methods, when the item difficulty level has equal probability of being answered among the subgroups, the item is considered as unbiased. Transforming item difficulty and variance analysis are among these methods.

Problem: Peabody Picture Vocabulary Test (PPVT) was used in many researches in order to find out whether there is a language difference among children according to gender and socio-economic level. The words used in this test were selected from the most frequently used words in American magazines and newspapers. Determination of the use of adapted test in the culture in which it is adapted according to different variables will provide detailed information on the reliability and validity of the test.

For this reason, PPVT items were analysed whether they include DIF in subgroups dependent upon socio-economic level and gender variables and the items displaying DIF were analysed from the point of item bias in this research. Answers of the following research questions are sought within the settings of this research:

1. Is there any differential item functioning in the items of PPVT from the points of gender and socio-economic level? Do the items displaying DIF include distractor bias?
2. Is the reliability of the test affected when the items displaying DIF exist in the test?
3. Do the methods of MH and LR used in detecting DIF give harmonious results? What are the similarities and differences of these two methods?
4. How can a DIF exists in an item be explained?

MATERIALS AND METHODS

592 children living in the city centre of Ankara form the sample of the research. 51% (300 children) of them are boys, 49% (292 children) of them are girls and 49% (295 children) of them are in high socio-economic level and 51% (297 children) of them are in low socio-economic level. The area where children have been living and their family income level were considered as criteria in determining socio-economic level.

Denver II Developmental Screening Test ("Turkey Standardization") was used in order to determine whether the children forming the sample are risky from the point of development and Peabody Picture Vocabulary Test was used in order to investigate the item bias.

PPVT was developed being applied to a group between the ages of 2-18 in America by Dunn in 1952 and was adapted to Turkish by being applied to children (n=1440) between the ages of 2-12 in 1974 by Katz, Önen, Demir, Uzlukaya and Uludağ [10]. The internal consistency (KR-20) of the scale varies between.71 and.81. Furthermore, test re-test reliability was found out between 0.52 and 0.90. The validity of similar scales carried out with Stanford-Binet and Wechsler varies between 0.71 and 0.91. The relevant scale has still been being used frequently in Turkey in taking educational decisions for children.

PPVT consists of 100 items and is scored as 1-0. Children are asked to find out the relevant picture which is related to the word stated in each item of the test. Item content consists of the pictures given in the alternatives rather than the words being asked.

In this research, items having variances equal to zero or near zero were not included in DIF analysis. Five items (items between 1-5) from the point of gender variable and 20 items (items between 1-10 and 91-100) from the point of socio-economic variable were not included in the analysis. The total numbers of the items included in DIF analysis were as follows respectively: 95 items for the gender variable and 80 items for the socio-economic variable.

Mantel-Haenszel and logistic regression methods were used in order to determine whether the items of PPVT displaying DIF. The items displaying DIF that were detected were analysed with distractor response analysis in order to find out whether there is bias in these items and also the origin of bias was sought. In the analyses, EZDIFF, ITEMAN, MS OFFICE (EXCEL) and SPSS 13.0 software packages were utilized.

Mantel-Haenszel: Mantel-Haenszel (MH) is a nonparametric method [11]. In this method, the scores of the groups were matched and comparisons were made between the people who have the same level of scores. In MH method, the level of attribute is kept as constant. Each difference in item performance between two groups displays differential item functioning (DIF).

α (odd ratio or MH-Alpha), measures the degree of performance difference in focal and reference groups. α values are interpreted as below:

If $\alpha = 1$, there is not any DIF.
If $\alpha < 1.00$? item includes bias against reference group.
If $\alpha > 1.00$,? item includes bias against focal group.

The degree of bias increases as long as α_becomes distant from 1.00 [12].

In MH method, DIF is determined in three levels [11, 13]:

Level A: If Δ-MH does not significantly differ from 0 and IΔ-MHI < 1, there is a DIF in the item in the level that can be neglected.

Level B: If Δ-MH significantly differs from 0 and 1 ≤ IΔ-MHI < 1.5, there is a DIF in the item in the medium level.

Level C: If IΔ-MHI ≥ 1.5, there is a DIF in the item in significant level.

Logistic Regression: Logistic regression (LR) is a regression model in which dependent variable can take two values and the independent variable is continuous variable. Logistic regression determines both the uniform DIF and nonuniform DIF. This situation is the functional aspect of logistic regression compared to other methods [11].

In order to calculate the effect size in logistic regression model, standardized regression coefficients are utilized. Standardized regression coefficients (R^2) display the degree of DIF and are determined in three levels [11,7,13]. Zumbo and Thomas [6] made a classification that can be calculated with the formula of $\Delta R^2 = R^2(M3)-R^2(M1)$ in determining DIF level. The classification that Zumbo and Thomas [6] made was given below:

Level A: If $\Delta R^2 < 0.13$, there is a DIF in the level that can be neglected.

Level B: If $0.13 \leq \Delta R^2 < 0.26$, there is a DIF in the moderate level.

Level C: If $\Delta R^2 \geq 0.26$, there is a DIF in significant level.

Distractor response analysis: This method covers the investigation of distractors within test items. Distractor function determines the significance degree of the difference between the frequencies of the responses given to the distractors by two or more groups. Distractor response analysis approach focuses only on the responses given and does not interested in item root. The responses left empty are omitted in the analysis. Distractor response analysis approach helps the people who prepared the test. It provides required point of view in selecting distractors. It provides the opportunity of controlling content validity [2].

Considering relevant literature, DIF in A level in an item can be neglected [11, 7, 13], therefore, distractor response analysis was performed only for the items in the levels of B and C in this research.

Findings: Whether there is a difference in the significant level from the points of score difference between groups was investigated by calculating the descriptive statistics obtained from the data. Descriptive statistics were given in Table 1.

When the data were grouped according to gender, the variance of the first five items was found out as very low and it obstructed these items being analysed with DIF determination methods. In the same way, when the data were grouped according to socio-economic level variable, the variances of the first 10 and the last 10 items were found out as very low and 20 items could not be included within the setting of the analyses.

Table 1: Descriptive Test Statistics According To Gender And Socio-Economic Level Variables

	Boys	Girls	High SEL	Low SEL
Item Number	95	95	80	80
Respondent Number	300	292	295	297
Mean	45.61	45.52	45.62	35.31
Standard Deviation	15.26	14.60	13.25	13.41
Reliability	0.94	0.94	0.93	0.93
T	0.083		-9.76	

** p<0,05

Table 2: Mantel-Haenszel Analysis Results of The Items Displaying DIF According To Gender Variable

Item NO.	α	χ^2	p	Δ-MH	DIF
17	1.980	4.952	0.026	-1.605	C
19	0.370	10.037	0.002	2.336	C
22	2.101	6.156	0.013	-1.744	C
27	1.995	9.569	0.002	-1.623	C
29	0.534	7.734	0.005	1.476	B
32	0.642	4.477	0.034	1.043	B
43	0.550	9.014	0.003	1.406	B
44	0.593	4.696	0.030	1.229	B
45	1.495	5.200	0.023	-0.946	A
49	0.532	10.495	0.001	1.483	B
54	1.694	6.785	0.009	-1.238	B
65	0.536	7.432	0.006	1.465	B
68	2.638	18.110	0.000	-2.279	C
71	1.709	4.704	0.030	-1.260	B
82	0.467	5.381	0.020	1.791	C
96	2.858	4.317	0.038	-2.468	C

Reference group: girls (n=292); Focal group: boys (n=300)

In the research, t test was used in order to test the difference between means in both groups and no significant difference was found out in gender subgroups in the level of α=,05. In other words, means of groups' test scores do not display difference in the significant level. On the other hand, when the socio-economic level was taken into consideration, difference in the significant level was found out between the two groups (p<0.05).

Findings obtained from the comparisons related to gender:

According to MH results from the point of gender variable, the items displaying DIF in various levels were as follows respectively; one item displayed DIF in level A, eight items displayed DIF in level B, seven items displayed DIF in level C out of 95 items and the total number of items displaying DIF was 16. Eight items displaying DIF out of these items were in the favour of reference group (girls) and the other eight items displaying DIF were in the favour of focal group (boys). MH analysis results of the items displaying DIF were given in Table 2.

Table 3: Logistic Regression Analysis Results of The Items Displaying DIF According To Gender Variable

Item No.	Group interaction p value	Group-characteristic Interaction p value	R^2	DIF Type	DIF Level
6	0.082	0.013	0.275	NU	C
13	0.025	0.033	0.279	U	C
22	0.039	0.234	0.414	U	C
28	0.014	0.016	0.507	U	C
31	0.100	0.025	0.229	NU	B
34	0.104	0.039	0.334	NU	C
39	0.041	0.039	0.466	NU	C
45	0.098	0.011	0.031	NU	A
63	0.053	0.031	0.307	NU	C
68	0.059	0.006	0.425	NU	C
72	0.072	0.043	0.314	NU	C
75	0.040	0.026	0.457	NU	C
96	0.013	0.005	0.239	NU	B
97	0.000	0.000	0.051	U	A

p<.05; U: Uniform, NU: non-uniform

According to logistic regression analysis results, the total number of the items displaying DIF was 14. Four of them displayed uniform DIF and 10 of them displayed nonuniform DIF. According to the classification used in determining the DIF level, the number of the items displaying DIF in LR analysis were as follows respectively; 2 items displayed DIF in level A, 2 items displayed DIF in level B and 10 items displayed DIF in level C. Logistic regression analysis results related to the items displaying DIF were given in Table 3.

DIF was found in various levels in 16 items with MH method and 14 items with LR method. The total number of the 27 items displaying DIF regarding methods was as follows: 41% of 27 items displayed differential item functioning according to MH method, 48% of them displayed differential item functioning according to LR method and 11% of them displayed differential item functioning according to both methods. It was observed that four of 27 items displayed DIF both with MH and LR methods (22, 45, 68 and 96. items). DIF level of 3 items (22, 45 and 68. items) which were found out displaying differential item functioning with both methods are same.

The relationship between two methods was found out weak in determining the items including DIF. It was also seen that the results of the methods used in determining DIF in other researches were not harmonious

Table 4: Internal consistency (KR-20) After The Removal Of Biased Items From The Test In Gender Subgroups

	Girls	Boys
Internal consistency (KR-20) when the biased items exist within the test	0.94	0.94
Internal consistency (KR-20) after the biased items were removed from the test	0.95	0.96

Table 5: Distractor Response Analysis Results Of The Items Displaying DIF According To Gender Subgroups

	Answer Choice			
Item No.	A χ^2	B χ^2	C χ^2	D χ^2
13	CA	4.36	1.97	4.03
17	CA	1.37	1.54	15.36*
19	3.3	1.49	CA	4.93
22	3.61	CA	0.88	3.73
27	0.41	CA	10.55*	6.53
28	0	CA	0	0.46
29	0.59	CA	4.03	2.17
31	0.53	0.11	2.96	CA
32	1.68	1.84	CA	1.63
34	CA	2.27	0.15	0.04
39	1.71	CA	0.39	2.64
43	CA	4.56	0.86	7.53*
44	0.07	0.69	CA	3.83
49	4.27	CA	7.92*	5.03
54	4.67	4.12	CA	0.93
63	CA	0.23	0.40	0.90
65	CA	3.78	7.99*	4.52
68	25.36*	1.11	9.73*	CA
71	CA	1.44	6.44	2.86
72	0.40	0.23	CA	0.85
75	1.31	2.79	0.05	CA
82	3.93	CA	0.18	6.70*

*p<0,05; α=1/4(.05)=0,0125 ; sd=1 ; χ^2=6,63; CA: Correct answer

with each other [11, 14-17]. Since the number of the sample is low, this increases the sensitivity of LR towards type 1 error. It was stated that LR will give more accurate results in large samples [18, 19].

Internal consistency (KR-20) was calculated when the items displaying DIF exist in test and after they were removed from the test in order to determine whether the items displaying DIF affect the reliability of the test. Results related to the internal consistency (KR-20) were given in Table 4.

The removal of items displaying DIF from the test increases the reliability of the test. Items removed could affect the internal consistency and reliability of the test in the negative direction. Rownozski and Reith [12] investigated the degree of the effects and impacts of the items displaying DIF in a test on measurement characteristics from the points of reliability and validity. They found out that the removal of biased items does not decrease the characteristics of measurement in the significant level. The replication of carrying out similar researches will make contributions to this field.

Responses given to the items including DIF and the situation of whether a bias existed dependent upon the distractors within these items were studied. The findings were given in Table 5.

When the items which were determined displaying DIF with MH method were studied, it was seen that there is bias related to the distractors in seven items. The distractors of the items displaying DIF and which were in favour of girls were analysed in order to find out whether there is a bias in favour of girls. It was seen that the distractors of four items affected the responses of boys in significant level and no difference was found out in the significant level from the point of response preference within the distractors of two items. One item (96 item) was discarded as a result of the low number of respondents. The distractors of the items displaying DIF and which were in favour of the boys were analysed and it was seen that that the distractors of three items affected the responses of girls in the significant level. No difference in the significant level was found out from the point of response preference in four items.

When the items including DIF according to logistic regression were studied, it was seen that the distractors of nine items did not affect the response preferences of any specific group in significant level. In one items, it was found out that there are biases dependent upon distractors. Two items (6, 96 and 97 items) were discarded from the analyses as a result of the low number of respondents.

Distractor biases were found out in seven items out of 16 items displaying DIF determined by MH method and in one items out of 14 items displaying DIF determined by LR method. When the results of MH and LR are taken into consideration with the distractor response analysis approach, we can think that MH method is much more sensitive towards the differences related to the distractors. It was stated that the best results were obtained from the MH test statistics in various researches in which the different methods are compared to each other in determining DIF [14-17].

Table 6: Mantel-Haenszel analysis results of the items displaying dif according to socio economic level variable

Item NO.	α	χ^2	p	Δ-MH	DIF Level
14	0.169	19.718	0.000	4.182	C
15	1.935	7.326	0.007	-1.552	C
19	2.208	5.967	0.015	-1.861	C
20	3.559	16.781	0.000	-2.983	C
22	2.006	4.667	0.031	-1.635	C
24	3.050	10.106	0.001	-2.621	C
27	1.671	5.151	0.023	-1.207	B
29	1.981	9.514	0.002	-1.607	C
30	3.831	39.296	0.000	-3.156	C
31	2.195	15.719	0.000	-1.847	C
32	6.222	76.114	0.000	-4.296	C
34	1.927	4.292	0.038	-1.542	C
35	3.738	42.933	0.000	-3.098	C
37	0.588	8.239	0.004	1.247	C
38	1.745	5.592	0.018	-1.308	C
39	2.280	5.994	0.014	-1.937	C
44	5.779	47.822	0.000	-4.123	C
45	1.935	13.297	0.000	-1.551	C
46	2.932	31.194	0.000	-2.528	C
47	2.003	11.942	0.001	-1.632	C
48	0.431	15.069	0.000	1.980	C
52	4.852	36.305	0.000	-3.712	C
53	3.017	24.111	0.000	-2.595	C
54	1.672	6.112	0.013	-1.208	B
55	1.990	12.617	0.000	-1.617	C
58	3.472	40.145	0.000	-2.925	C
60	1.859	8.644	0.003	-1.456	B
61	0.641	4.523	0.033	1.047	B
62	2.663	23.513	0.000	-2.301	C
66	2.476	17.150	0.000	-2.130	C
74	0.643	3.993	0.046	1.038	B
86	0.422	5.858	0.016	2.028	C
87	0.361	6.352	0.012	2.397	C

p < 0,05; Reference group: children in high socio-economic group (n=295); Focal grup: children in low socio-economic group (n=297)

Jensen [1] made a similar research on white, black and Mexican-American children utilizing PPVT. Statistically significance was not found among white girls, black boys and black girls. Statistically significance was found out in favour of Mexican-American girls and boys compared to others (p<.01). Distractor response analysis was used in this research and a difference in the ratio of 26% was found out between the distractor preferences of black and white children.

Findings obtained as a result of comparing socio-economic level:

Table 7: Logistic regression analysis results of the items including dif according to socio-economic level variable

Item No.	Group interaction p value	Group-characteristic Interaction p value	R^2	DIF Type	DIF Level
11	0.000	0.000	0.244	U	B
12	0.004	0.004	0.408	NU	C
13	0.003	0.000	0.262	NU	C
14	0.020	0.714	0.262	U	C
15	0.112	0.013	0.308	NU	C
16	0.175	0.020	0.311	NU	C
20	0.014	0.214	0.212	U	B
23	0.047	0.035	0.078	NU	A
27	0.076	0.009	0.225	NU	B
32	0.999	0.012	0.376	NU	C
35	0.758	0.025	0.328	NU	C
36	0.023	0.067	0.361	U	C
37	0.002	0.036	0.087	U	A
44	0.948	0.032	0.408	NU	C
45	0.033	0.263	0.024	U	A
46	0.016	0.441	0.217	U	B
66	0.045	0.008	0.528	NU	C
67	0.003	0.002	0.180	NU	B
79	0.101	0.042	0.375	NU	C
83	0.015	0.025	0.331	U	C

p<0,05 ; U: Uniform, NU: non-uniform

According to Mantel-Haenszel results, DIF was found in five items in level B, in 28 items in level C and the total number of items displaying DIF is 33. 25 items displaying DIF were found out in the favour of reference group (high socio-economic group) and three items displaying DIF were found out in the favour of focal group (low socio-economic group). MH analysis results displaying DIF were given in Table 6.

According to logistic regression analysis results, the total number of the items displaying DIF according to socio-economic group variable was 20. There were eight uniform DIF and 12 nonuniform DIF within these items. According to the classification used in determining DIF level, DIF was found out in three items in level A, five items in level B and 12 items in level C in LR analysis. LR analysis results of the items displaying DIF were given in Table 7.

DIF was found out in 33 items with MH method and in 20 items with LR method in various levels, therefore, the overall number of the items displaying DIF was 42. Differential item functioning was found out in 11 items out of these items with both MH and LR methods. 52,3% of 42 items displayed DIF according

Table 8: Internal consistency (KR-20) after the removal of biased items from the test in low and high socio economic level subgroups

	Low SEL	High SEL
Internal consistency (KR-20) when the biased items exist within the test	0.93	0.93
Internal consistency (KR-20) after the biased items were removed from the test	0.97	0.97

to MH method, 21.4% of them displayed DIF according to LR method and 26.1% of them displayed DIF according to both methods.

When the data related to the children in high and low socio-economic groups were analysed, it was seen that there was not an exact harmony between the two methods in determining the items displaying DIF. When the analysis was considered from socio economic level variable, it was previously mentioned that both MH and LR methods provided common results in 11 items. It was seen that both methods were in harmonious state in determining DIF level. DIF level of all items were found out same in two methods except four items (20, 37, 45 and 46 items). However, DIF level of methods displayed differences in four items.

Whether the items displaying DIF affect the internal consistency (KR-20) is an important matter that should be considered. Internal consistency (KR-20) was calculated regarding both the items displaying DIF exist in the test and after they were removed from the test. Data related to internal consistency (KR-20) were given in Table 8.

Removal of the items from test increases internal consistency (KR-20). These items could affect the internal consistency coefficients (KR-20) and validity of the test in the negative direction.

When the smallness of test items' variance is considered as one of the factors that increases the internal consistency of the test [20, 21], it can be stated that DIF spoils the homogeneity of the test items. Since DIF existing in an item increases the variances of the test items, this decreases the internal consistency of the test. Removal of the items displaying DIF can make a contribution in the increase of the reliability which means the internal consistency of the test. Therefore, the effect of DIF on the internal consistency of the test was also studied in this research.

Statistically significant difference ($p<0.05$) was found out between the score means of two groups in the socio-economic level. When the difference between score means of two groups was considered after removal of the items displaying DIF, it was seen that the difference between the score means decreased (-6,60). It was seen that the items including DIF increases the

Table 9: Distractor response analysis results of the items displaying dif according to low and high socio economic level groups

Item	Answer Choice			
No	A χ^2	B χ^2	C χ^2	D χ^2
11	10.5*	0.64	CA	2.57
12	0.01	CA	1.84	5.68
13	CA	0	2.05	4.08
14	CA	0.92	4.98	0.12
15	6.74*	26.3*	CA	3.16
16	6.65*	2.17	8.01*	CA
19	6.39	10.1*	CA	7.04*
20	4.05	12.2*	21.7*	CA
22	13.8*	CA	1.4	13.1*
24	CA	10.2*	1.53	18.9*
27	28.3*	CA	0.64	2.11
29	14.6*	CA	14.3*	15.9*
30	20.5*	43.6*	11.7*	CA
31	21.1*	12.9*	7.73*	CA
32	62.4*	35.9*	CA	61.7*
34	CA	4.81	4.22	4.94
35	74.3*	8.54*	15.4*	CA
36	CA	0.53	1.12	5.48
38	3.35	CA	10	5.38
39	8.32*	CA	1.05	6.16
44	48.5*	15.8*	CA	21.2*
46	9.37*	38.2*	7.75*	CA
47	20.7*	CA	20.4*	2.94
48	2.38	CA	7.06*	0.85
52	14.2*	CA	16.8*	29
53	14.4*	15.6*	9.95*	CA
54	3.36	8.18*	CA	7.46*
55	12.4*	CA	9.99*	4.12
58	34	55.3*	7.84*	CA
60	CA	19.7*	11.6*	3.95
61	CA	4.18	9.35*	20.5*
62	33	16.7*	CA	15.2*
66	CA	28.7*	14.3*	15.4*
67	5.07	CA	2.14	0.15
74	CA	0.18	0.22	3.79
79	0.02	3.31	CA	1.63
83	0.94	CA	0.06	0.01
86	1.97	0.58	1.71	CA
87	8.26*	1.33	CA	0.22

*p<0,05; α=1/4(.05)=0,0125 ; sd=1 ; χ^2=6,63; CA: Correct answer; HSEL: high socio economic group, LSEL: low socio economic group

difference between score means. Daha önce de belirtildiği gibi, Ackerman [4] stated that the difference in the performance in measured variable of two groups can cause a bias. Since the performance in the items related to groups' measured variable differs, it is possible to find DIF in an item. Since the difference in groups' performance increases when the items including DIF are removed from the test, it can be said that the difference between the two groups' score averages decreases.

Responses related to the items displaying DIF and whether there are biases within these responses were analysed. Results were given in Table 9:

As a result of distractor response analyses, it was seen that some distractors in 13 items affected response preferences of low socio-economic group in significant level and all distractors in 11 items affected response preferences of low socio-economic group in significant level. When the distractors of the items that are in the favour of children in low socio-economic group were studied, no difference in the significant level was found in the distractors of two items from the point of response preference. Distractor A in 87. item affected the response preferences of children in high socio-economic group in the significant level.

When LR and distractor response analyses results were studied together, no difference in the significant level was found in seven items from the point of response preference, on the other hand, biases related to distractors were found in 10 items. It was found out that there were distractor biases in 26 items out of 31 items displaying DIF determined by MH method and in 10 items out of 17 items displaying DIF determined by LR method. When the results of MH and LR were studied considering the distractor response analysis, we can think that MH method is much more sensitive to the differences dependent upon distractors.

Content analyses of items displaying DIF from the points of socio economic level and gender variables were done being investigated by subject field specialists and the reasons why the items displayed DIF were investigated. The reasons of why the items displayed DIF were determined as below:

- Item pictures
- Differences related to experiences and interests
- Method
- External bias

RESULTS AND DISCUSSION

In the research, interviews were done with 592 children in 4-6 age groups, Peabody Picture Vocabulary Test and Denver II Developmental Screening Test were used. Data were grouped considering socio-economic level and gender variables and analysed with MH and LR methods. Results can be summarized considering the data obtained as follows:

About 27% of test items displayed DIF according to gender and it was seen that these items decreased the reliability of the test.

About 52% of test items displayed DIF according to socio-economic level and it was seen that these items decreased the reliability of the test. Besides, when the items displaying DIF were removed from the test, it was seen that the difference between score means of two groups decreased but it did not disappear completely. It was found out much more items displaying DIF and including distractor bias in the comparisons done considering socio-economic level variable.

Besides, when the biased items are removed from the test, it is seen that the internal consistency of the test increases. This increase becomes higher when the biased items related to socio-economic level variable are removed. Findings revealed that the biased items affect the reliability of the test in small amount.

No harmonious result could be obtained between two methods. It was found out that MH method was much more sensitive to distractor bias.

The origin of DIF was investigated in the items displaying DIF and it was seen that there were four reasons. The reasons that can form DIF are classified as: pictures (content), differences related to experiences and interests, method and external bias. Within the findings obtained in this research, items displaying DIF of Peabody Picture Vocabulary Test should either be removed from the test or changed considering revisions and it should be useful to obtain data related to the functions of the items in various groups as a result of changes done. Studying the test sessions in different cultural groups can provide data whether the items of the test dependent upon culture. Besides, the external bias of Peabody Picture Vocabulary Test can be determined by taking language test as an external criterion and the effect of items displaying DIF on internal consistency (KR-20) can be determined.

REFERENCES

1. Jensen, A.R., 1980. Bias in mental testing. Methuen.
2. Osterlind, S.J., 1983. Test item bias. Sage Publication.
3. Embretson, S.E. and S.P. Reise, 2000. Item response theory for psychologists. Lawrence Erlbaum Associates, Publishers, London.
4. Ackerman, T.A., 1992. A didactic explanation item bias, item impact and item validity form a multidimensional perspective. Journal of Educational Measurement, 29(1): 67-91.
5. van de Vijver, F.J.R. and Y.H. Poortinga, 2005. Conceptual anda methodogical issues in adapting testing. In R.K. Hambelton, P.F. Merenda and C.D. Spielberger (Ed.), Adapting educational and psychological tests for cross-cultural assessment. pp: 39-63. London.
6. Zumbo, B.D. and D.R. Thomas, 1997. A measure of effect size for a model-based approach for studying DIF. Prince George, Canada: University of Northern British Colombia.
7. Zumbo, B.D., 1999. A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modelling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, On: Directorate of Human Resources Research and Evaluation, Department of: Nationale Defense, Canada.
8. Hambleton, R.K., 2005. Issues, designs and technical guidelines for adapting test into multiple languages and cultures. In R. K. Hambelton, P.F. Merenda and C. D. Spielberger (Ed.), Adapting educational and psychological tests for cross-cultural assessment. pp: 3-38. London.
9. Allalouf, A., 2003. Revising translated differential item functioning items as a tool for improving cross-lingual assessment. Applied Measurement in Education, 16(1): 55-73.
10. Özgüven, İ.E., 2004. Psikolojik testler (Psychologic tests). PDREM Publications, Ankara.
11. Gierl, M., S.N. Khaliq and K. Boughton, 1999. Gender differential item functioning in mathematics and science: Prevalence and policy implications. Paper Presented at the Symposium Entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education. Canada, June.
12. Roznowski, M. and J. Reith, 1999. Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement?. Educational and Psychological Measurement, 59(2): 248-269.
13. Hidalgo, M.D. and J.A. López-Pina, 2004. Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. Educational and Psychological Measurement, 64(6): 903-915.
14. Gomez-Benito, J. and J. Navas-Ara, 2000. A comparison of χ^2, RFA and IRT based procedures in the detection of DIF. Springer Netherlands. 34(1): 17-31.
15. Gondal, M.B., 2001. Differential item functioning analysis of 4-graders' science and Urdu (national language) achievement test items in Pakistan. Unpublished PhD Thesis Middle East Technical University Institute of Social Sciences, Ankara.
16. Bertand, R. and N. Boiteau, 2003. Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural contexts using TIMSS data. EDRS Price.
17. Emenogu, B. and R.A. Childs, 2003. Curriculum and translation DIF: A comparison of two DIF detection techniques. Canada Ontario.
18. Pang, X., F. Tian and M.W. Boss, 1994. Performance of Mantel-Haenszel and logistic regression DIF procedures over replications using real data. American Educational Research Association. April 4-8.
19. Jodoin, M.G. and M.J. Gierl, 2001. Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. Applied Measurement in Education. 14: 329-349.
20. Crocker, L. and J. Algina, 1986. Introduction to classical and modern theory. CBS College Publishing.
21. Baykul, Y., 2000. Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulamasi (Measurement in education and psychology: Classical test theory and its application). ÖSYM Publications, Ankara.