European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.1199695 Volume 4 | Issue 3 | 2018

HOW THE ORDER OF THE ITEMS IN A BOOKLET AFFECTS ITEM FUNCTIONING: EMPRICAL FINDINGS FROM COURSE-LEVEL DATA?

Akif Avcu[i],
Emine Burcu Tunç,
Müge Uluman
Marmara University, Atatürk Eğitim Faculty,
Educational Sciences Department, Turkey

Abstract:
In many test applications, from larger to smaller scale, items could be given in different booklets with different ordering. Differing the place of items may cause items to function differently. The effect of item order is defined as the interaction between response of a test taker on any item and the order of the item on the test. When earlier studies examined, it can be seen that the item order effect shouldn't not be ignored. In this study, it was aimed to examine whether items included in graduate level Measurement and Evaluation test showed differential item functioning (DIF) between different booklets. Booklet effect was investigated by a series of DIF detection methods and analysis were carried out using R program. The data collected with 25 multiple-choice item test. The test covers the curriculum of the undergraduate level Measurement and Evaluation course. When ordering the items in booklet, the sequence of curriculum was followed on one booklet (Booklet A) while, it is not the case for the second booklet (Booklet B). The results showed that the order of the items have an effect on DIF for some items. It was concluded that, item location may affect the probability of respondents to correctly answer to items.

Keywords: test booklet effect, Differential Item Functioning

1. Introduction

Differential Item Functioning (DIF) is a statistical technique of matching individuals according to their abilities in relation to the variable to be measured, and then detecting those individuals in different groups possessing different possibilities of responding to the item (Camilli and Shepard, 1994; Zumbo, 1999). As developed by the Educational Test Service (ETS) in 1986, DIF is a preliminary step to determine item bias, and has

[i] Correspondence: email avcuakif@gmail.com

become a standard in bias analysis (Roever, 2005). There are a number of studies conducted within the scope of DIF, but there is no consensus reached by experts on the source of DIF (Karami and Nodoushan, 2011).

In the studies conducted so far, translational problems and cultural differences (Asil, 2010; Girl and Khaliq, 2001), the format (open ended vs multiple choice) of the items (Feng, 2008; Henderson, 1999; Qian, 2011; Zenisky, Hambleton and Robin, 2003), different scoring models (Gelin and Zumbo, 2003; Henderson, 2001; Tunç, 2016) and item contents (Liu and Wilson, 2009; Mendes-Barnett and Ercikan, 2006; Ong et al., 2011) are shown as possible sources of DIF. Another variable considered to have an effect on is the order of item in booklet.

Some of the considerations that are important for item order is placing the the items from the easy to the difficult and placing the items belonging to the same subject are given together (Bahar, Nartgün, Durmuş and Bıçak, 2010; Güler, 2011; Özçelik, 2010; Tekin, 1993). However, these principles are often overlooked. For this reason, when test items placed in different orders without obeying aforementioned principles, they could be perceived more difficult or easier and they could show bias between the booklets.

The effect of item order is defined as the interaction between response of a test taker on any item and the order of the item on the test (Kingston and Dorans, 1984; Qian, 2014). When earlier studies examined, it can be seen that the item order effect shouldn't be ignored. Hambleton (1968) obtained two different test booklets by sorting the items from "easy to the hard" and "hard to easy", and stated that the average score of the students is lower for the booklet in which the items are arranged in order of "hard to easy". Newman, Kundert, Lane and Bull (1988) reached a similar conclusion and stated that when cognitive difficulty levels of items are arranged from "easy to hard" test takers tend to be more successful. Tippets and Benson (1989), who study the relationship between item ordering in different booklets and anxiety level of students, stated that item order may affect students' level of anxiety. Yergin (2007) investigated how item orders affect the success of students and suggested the use of "easy to hard" item ordering when ordering the items in booklets.

When studies related to the effect of item order on DIF were examined, it was concluded that the order of items is cause of DIF (Abedi, Leon and Kao, 2007; Freedle and Kostin, 1991; Schmitt and Dorans, 1988). Similarly, Erdem (2015), Yilmaz (2014), Yakar and Yavuz (2014) have shown that the number of items functioning differently could change according to different test booklets. As become seen, even the effect of item order based on difficulty of items are well studied, the effect of curriculum based ordering is not investigated well.

In this study, it was aimed to examine whether items included in graduate level Measurement and Evaluation test showed DIF according to different test booklets. Despite the frequent use of different test booklets in large scale test applications and related literature, the booklet-based DIF studies in small university test applications don't take part in this intense literature. For this reason, it is considered that this study is important and that the results obtained will be informative for similar applications.

2. Methodology

2.1 Data Source

The data collection tool consists of a total of 25 multiple-choice items with five choices. The test covers the curriculum of the undergraduate level Measurement and Evaluation course. The exam was applied in an education faculty of state owned university located in Istanbul. The items were presented to the students in two different booklets. Booklet A (reference group) was prepared according to the order of the course curriculum topics during the semester. Item ordering in Booklet B (focus group) was prepared as described below.

The first five items in Booklet A were included as the last five items in Booklet B. In the same way, the last five items of Booklet A are listed as the first five items in Booklet B. The order of all other items (from 6. item through 20. item) remained the same in both booklets. In this way, item ordering is fixed similarly for Booklet A, while students who took Booklet B started to the exam with items covering the most recent topics thought in the class. The order of the items in the booklet and item difficulties are presented in Table 2. To keep other factors constant, the correct answer options stayed the same for both booklets that are Booklets were randomly distributed to the students who took the test and each test Booklet was received by 425 students.

2.2 Data Analysis

Analysis of the data was carried out using R program (R Development Core Team, 2008). For this analysis, the "*difR*" package, developed by Magis et al. (2010), was used. The results for Mantel-Haenszel, Lord's Chi-Square and Logistic Regression methods were provided in the study.

A. Mantel-Haenszel

The Mantel-Haenszel (MH) method is a statistical method in which contingency tables are used. This method (Mantel and Haenszel, 1959) aims to test whether there is a relationship between group membership and item response depending on the total test score. The number of test items is J, the number of test items is T_j, the total score is taken as j (j will take a value between zero and J). Then, for any item tested, the test takers are classified for the 2X2 probability table with group membership and response type (true or false) inputs. A_j, B_j, C_j and $D_j$are the values obtained for the four cells of this table, and A_j and B_j represent the numbers of correct and incorrect answers of the test item in the reference group respectively. Similarly, C_j and D_j correspond to the corresponding numbers of true and false answers in the focus group. The relevant table is given in Table 1 below.

Table 1: Mantel-Haenszel Table

Group	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{rj}
Total	M_{1j}	M_{0j}	T_1,

Depending on these values, the MH statistic is calculated as:

$$MH\chi^2 = \frac{\left[\left|\sum_j A_j - \sum_j E(A_j)\right| - .5\right]^2}{\sum_j Var(A_j)}$$

In the equation (1), -.5 correction is the continuity correction factor used to improve the prediction of the chi-square distribution, which is particularly necessary for small frequencies.

B. Lord's Ki-Square Statistics

The Lord's (1980) chi-square method tests the differences in one or more item parameters at the same time between reference and focus groups. The difference vectors in the item parameters and the inverse of the variance-covariance matrix for these differences are used to calculate a chi-square statistic that is compared to a critical value based on a predefined significance level. The degree of freedom (df) used in testing the significance corresponds to the number of parameters examined for each item. The generated difference vector is expressed as:

$$V= (\hat{a}_F - \hat{a}_R , \hat{b}_F - \hat{b}_R)$$

The test statistic is expressed as:

$$Q = VS^{-1}V$$

Where S is the variance-covariance matrix of the differences between the item parameters. Q follows a chi-square distribution with a degree of freedom equal to the estimated number of parameters (Lord, 1980).

C. Logistics Regression

The logistic regression method uses the group variable (1 = reference, 2 = focus), the total scale score for each item (expressed as variable TOT), the group & TOT interaction and total scale score for a group as independent variables while it uses response to any item (0 or 1) as dependent variable. It can be said that the Logistics Regression is a bridge between IRT and CTT based methods (Camilli and Shepard, 1994).

The logistic regression equation is:

$Y = b_0 + b_1TOT + b_2FORM + b_3TOT*FORM.$

Where Y is a natural logarithm of the probability ratio. So, the equation is:

$$\text{In}\left[\frac{p_i}{1-p_i}\right] = b_0 + b_1 tot + b_2 group + b_3 tot{*}group$$

Where p is the proportion of individuals contributing the item in the latent variable direction. DIF can then be examined by Chi-square test, taking as the degree of freedom 2 for both uniform and non-uniform DIF.

3. Findings

Before the findings related to DIF are given, the difficulty indices of the items in Booklet A and B in Table 2 are presented.

Table 2: Item orders and difficulties for Booklet A and B

Booklet A		Booklet B		
Item order	**Item difficulty**	**Item Order****	**Item Difficulty**	**Difference***
1	0.28	1 (21)	0.09	-0.19
2	0.36	2 (22)	0.41	**0.05**
3	-0.46	3 (23)	-0.72	-0.26
4	-1.77	4 (24)	-2.26	-0.49
5	2.74	5 (25)	2.39	-0.35
6	-2.05	6	-1.64	**0.41**
7	-0.49	7	-0.49	0
8	0.66	8	0.56	-0.1
9	-1.27	9	-1.57	-0.3
10	1.05	10	1.20	**0.15**
11	-0.02	11	-0.38	-0.36
12	-1.94	12	-2.04	-0.1
13	-0.57	13	-0.52	**0.05**
14	-0.01	14	-0.04	-0.03
15	0.84	15	0.81	-0.03
16	-2.03	16	-1.70	**0.33**
17	-0.32	17	-0.45	-0.13
18	-1.11	18	-1.12	-0.01
19	0.20	19	-0.15	-0.35
20	0.62	20	0.39	-0.23
21	-1.40	21 (1)	-1.16	**0.24**
22	3.93	22 (2)	4.40	**0.47**
23	-0.32	23 (3)	-0.27	**0.05**
24	-1.03	24 (4)	-0.95	**0.08**
25	-1.31	25 (5)	-1.30	**0.01**

* The items shown in bold in the difference column are the items that are identified as more difficult in form B than in form A. ** Values in brackets for Booklet B correspond to the order of item in Booklet A.

Estimation of the difficulty indices was performed using the Rasch model based on the IRT approach using the "*ltm*" package developed by Rizopoulos (2006) using the

R program. Item sequences in the table are arranged according to Booklet A. The order of items in Booklet B is shown in parentheses in the order of the original form.

As stated earlier, the last five items in booklet B corresponds to first five items in Booklet A. Four of these items had lower difficulty indices in Booklet B than in Booklet A, and one had a higher value. The first five items in Booklet B correspond to the last five items in Booklet A. All of these items had higher difficulty indices in Booklet B compared to Booklet A. When all the items are examined, 14 items in Booklet B have lower difficulty indices than Booklet A, 1 item is the almost the same and 10 items have higher difficulty indices.

The data obtained from the Booklets were analyzed according to the MH method and the results are shown in Table 3 whether the materials show uniform DMF according to different Booklet types.

Table 3: Result of Mantel-Haenszel Chi-square statistic

İtems	Statistic	P - value
1	0.28	0.60
2	1.13	0.29
3	1.51	0.22
4	1.83	0.18
5	0.58	0.45
6	4.89	0.03**
7	0.17	0.68
8	0.00	0.98
9	2.26	0.13
10	0.30	0.58
11	3.64	0.06*
12	0.59	0.44
13	0.00	0.94
14	0.23	0.63
15	0.65	0.42
16	2.53	0.11
17	0.25	0.61
18	0.01	0.94
19	2.11	0.15
20	0.08	0.77
21	3.23	0.07*
22	1.19	0.28
23	0.65	0.42
24	0.03	0.86
25	0.02	0.88

**p<0.05, *p<0.1

In all iterations, it was found that 24 items didn't show DIF at 05 significance level. As a result of the convergence, item 6, which is the only item showing DIF according to Table 3. The two items (items 11 and 21) showed a DIF at 0.1 signifigance lecel, but were not considered as items with DIF because this value was above the threshold value (3.84).

In Table 4, the magnitudes of impacts are included in the interpretation of three categories of MH method. In the second column of the table, the common odds ratios called 'AlphaMH' are estimated; in the third column are the effect sizes called 'DeltaMH' and in the last column are given the ETS Delta scale classification. According to this classification, 'A' is categorized as negligible effect, 'B' as intermediate effect and 'C' as big effect.

Table 4: DIF effect sizes for MH method

İtems	AlphaMH	DeltaMH	ETS Delta ölçeği
1	1.10	-0.23	A
2	0.84	0.41	A
3	1.22	-0.48	A
4	1.31	-0.63	A
5	1.19	-0.41	A
6	0.64	1.03	B
7	0.93	0.18	A
8	0.98	0.04	A
9	1.30	-0.62	A
10	0.91	0.22	A
11	1.34	-0.69	A
12	1.17	-0.37	A
13	0.98	0.05	A
14	0.92	0.21	A
15	0.87	0.33	A
16	0.73	0.74	A
17	1.09	-0.21	A
18	0.97	0.07	A
19	1.25	-0.53	A
20	1.05	-0.12	A
21	0.74	0.72	A
22	0.66	0.99	A
23	0.87	0.33	A
24	0.96	0.10	A
25	1.04	-0.09	A

Effect size codes: 0 'A' 1.0 'B' 1.5 'C'

As seen in Table 4, DIF is seen only in item 6 according to different test forms. As can be seen in Table 2, item 6 appears to be a more difficult item in Booklet B. The strength of the item in A Booklet is -2.05 while it is -1.64 in form B. The difference is 0.41, and students who take Booklet A having advantages over the ones who get Booklet B. The results were also provided in Figure 1. While only the 6th item is shown in red above the threshold value, the 11th item is almost at the threshold and the 21th item is just below threshold.

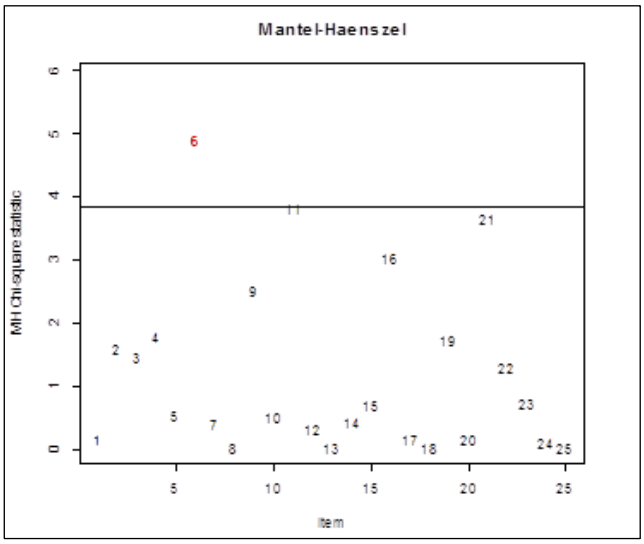

Figure 1: MH statistics and critical value

The results were also obtained using different DIF methods and are given in Table 5.

Table 5: DIF results for other methods

Items	Lord	M-H	Logistic	#DIF
1	NoDIF	NoDIF	NoDIF	0/3
2	NoDIF	NoDIF	NoDIF	0/3
3	NoDIF	NoDIF	NoDIF	0/3
4	NoDIF	NoDIF	NoDIF	0/3
5	NoDIF	NoDIF	NoDIF	0/3
6	***NoDIF***	***DIF***	***DIF***	***2/3***
7	NoDIF	NoDIF	NoDIF	0/3
8	NoDIF	NoDIF	NoDIF	0/3
9	NoDIF	NoDIF	NoDIF	0/3
10	NoDIF	NoDIF	NoDIF	0/3
11	*DIF*	*NoDIF*	*NoDIF*	*1/3*
12	NoDIF	NoDIF	NoDIF	0/3
13	NoDIF	NoDIF	NoDIF	0/3
14	NoDIF	NoDIF	NoDIF	0/3
15	NoDIF	NoDIF	NoDIF	0/3
16	NoDIF	NoDIF	NoDIF	0/3
17	*NoDIF*	*NoDIF*	*DIF*	*1/3*
18	NoDIF	NoDIF	NoDIF	0/3
19	NoDIF	NoDIF	NoDIF	0/3
20	NoDIF	NoDIF	NoDIF	0/3
21	*NoDIF*	*NoDIF*	*DIF*	*1/3*
22	NoDIF	NoDIF	NoDIF	0/3
23	NoDIF	NoDIF	NoDIF	0/3
24	NoDIF	NoDIF	NoDIF	0/3
25	NoDIF	NoDIF	NoDIF	0/3

In Table 5, the results for three different DIF detection methods are shown. According to these results, DIF was not captured in 21 items with none of the methods. Items 11

and 17 were found to have higher difficulty parameter values on Booklet A. Item 6 was detected as DIF showing item by two methods while remaining three methods was captured by only Logistic Regression Method. Twenty first item is the only one that its location changed across the forms. The location of the remaining three items is the same on both forms.

Item 21, which is captured as DIF, is related to topic of the properties of measuring instruments. Item 21 of the Booklet A is listed as the first item in the Booklet B. Therefore, the students who took book B did not start the test in parallel with the curriculum. According to the results of the analysis, item 21 in item B have higher difficulty level than item A in its booklet. As a reason for the observed DIF in this direction, the order of the differentiating items in the forms can be shown. In parallell with the results of MH method, DIF was captured according to the Logistic method for sixth item. It is an important item for research because it is located immediately behind the changing bunch of five items. Interestingly it is the easiest item in the test. While the items in the A Booklet were in an array parallel to the subject order, this was not the case for B Booklet. So, a student who takes Booklet B, starts the test with questions measuring the most recent topics covered, cannot easily found the correct response. The difficulty of this item for students who took the form B and the emergence of DIF in this item proved that the item order affected the DIF.

4. Discussion

Item order in a test is very important for test development. It has also been proven through studies that the effect of item ordering has an effect on the performance of individuals. In the current study, item ordering effects are examined and it was concluded that the order of the item in different booklets is effective on DIF. Similarly, Erdem (2015), Yilmaz (2014), Yakar and Yavuz (2014) have shown that the number of DIF detected items can change according to different test forms. However, these studies focus on different DIF detection sources (e.g. gender) over different forms. In the current study, booklet type was considered as a grouping variable and it was determined that DIF appeared according to different Booklets.

In this study, the items on the test forms were not ordered by their difficulty. In order to be able to accomplish this, item difficulties must be known prior, which is generally the case for large-scale test applications. In this present study, the ordering of the items was carried out by considering the subject line of the curriculum. In small scale exams like the ones applied at universities, items are generally ordered by considering the sequence of topics on curriculum.

Although booklet based DIF studies were made from the test forms in the large-scale exams, it was seen that there is no study for the smaller-scale test applications in which the sequence of items is generally ordered according to the subject matter in the curriculum. With this respect, it was thought that, findings from the current study has value when considering smaller test applications. As a recommendation, the same study can be repeated considering item order based on item difficulty indices. In

addition, bias determination can be carried out in accordance with the expert opinion for the specified DIF detected items.

References

1. Abedi J. Leon S. & Kao J. (2007). *Examining differential item functioning in reading assessments for students with disabilities.* Minneapolis. MN: University of Minnesota. Partnership for Accessible Reading Assessment.
2. Asil M. (2010). *Uluslararası Öğrenci Değerlendirme Programı (PISA) 2006 öğrenci anketinin kültürler arası eşdeğerliğinin incelenmesi*. Doktora Tezi. Hacettepe Üniversitesi. Ankara.
3. Bahar M. Nartgün Z. Durmuş S. & Bıçak B. (2010). *Geleneksel-tamamlayıcı ölçme ve değerlendirme teknikleri öğretmen el kitabı.* Ankara: Pegem Akademi.
4. Camilli G. and Shepard L. A. (1994). *Methods for identifying biased test items* (volume 4). California: SAGE Publications. Inc.
5. D. Magis S. Beland F. Tuerlinckx P. De Boeck (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research* Methods. 42. 847-862.
6. Erdem B. (2015). *Ortaöğretime geçişte kullanılan ortak sınavların değişen madde fonksiyonu açısından kitapçık türlerine göre farklı yöntemlerle incelenmesi.* Yüksek Lisans Tezi. Hacettepe Üniversitesi. Ankara.
7. Feng Y. (2008). *Difference in gender differential item functioning patterns across item format and subject area on diploma examinations after change in administration procedure.* Unpublished doctor dissertation. University of Alberta. Edmonton. Alberta. Canada.
8. Freedle R. & Kostin I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages (ETS Research Report RR-91-29). Princeton. NJ: Educational Testing Service.
9. Gelin M. N. & Zumbo B. D. (2003). DIF results may change depending on how an item is scored: An illustration with the Center for Epidemiological Studie Depression (CES-D) Scale. *Educational and Psychological Measurement.* 63. 65.
10. Girl M.J. & Khaliq S.N. (2001) Identifying sources of differential item and bundle functioning on translated achievement tests: A Confirmatory Analysis. *Journal of Educational Measurement;* Vol. 38. No. 2.; pp. 164-187.
11. Güler N. (2011). *Eğitimde ölçme ve değerlendirme.* Ankara: Pegem Akademi.
12. Hambleton R. K. (1968). *The effects of item order and anxiety on test performance and stress*. Paper presented at the meeting of American Educational Research Association. Chicago. Illinois. America.
13. Henderson D. L. (2001). *Prevalence of gender DIF in mixed format high school exit examinations.* Paper presented at the Annual Meeting of the American Educational Research Association. Seattle.

14. Henderson D. L. (1999). *Investigation of differential item functioning in exit Examinations across item format and subject area*. Unpublished doctor dissertation. University of Alberta. Edmonton. Alberta. Canada.
15. Karami H. and Nodoushan M. A. S. (2011). Differential item functioning (DIF): current problems and future directions. *International Journal of Language Studies. 5-4: 133-142.*
16. Karasar, N. (2005) *Bilimsel Araştırma Yöntemi*, Ankara: Nobel.
17. Kingston N. M. & Dorans N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*. 8. 147–154.
18. Liu L. O. & Wilson M. (2009). Gender Differences in Large-Scale Math Assessments. PISA Trend 2000 and 2003. *Applied Measurement in Education*. 22:2. 164- 184.
19. Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (s. 181-223). Hillsdale NJ: Erlbaum.
20. Mantel, N., Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 22: 719–748.
21. Mendes-Barnett S. ve Ercikan K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education. 19*. 289-304.
22. Nevman D. Kundert D. K. Lane D. S. & Bull. S. (1988). Effect of varying item order on multiple-choice test scores: importance of statistical and cognitive difficulty. *Applied Measurement in Education*. 1 (1). 89-97.
23. Ong Y.M. Williams. J. S. and Lamprianou I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing. 11*. 271-293.
24. Özçelik D. A. (2010b). *Test hazırlama kılavuzu*. Ankara: Pegem Akademi.
25. Qian J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Measurement in Education.* 38(7). 518–534.
26. R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
27. Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17 (5), 1-25. URL http://www.jstatsoft.org/v17/i05/
28. Roever C. (2005). That's not fair! Fairness. bias. and differential item functioning in language testing. *SLS Brownbag*. 9(15). 1-14.
29. Schmitt A. P. & Dorans N. J. (1988). Differential item functioning for minority examinees on the SAT. Princeton. NJ: Educational Testing Service.
30. Tekin H. (1993). *Eğitimde ölçme ve değerlendirme.* Ankara: Yargı Kitap ve Yayınevi.
31. Tippets E. & Benson J. (1989). The effect of item arrangement on test anxiety. *Applied Measurement in Education.* 2(24). 289-296.

32. Tunç E. B. (2016). *İki ve çok kategorili puanlanan maddelerde değişen madde fonksiyonlarının karşılaştırılması.* Doktora Tezi. Ankara Üniversitesi. Ankara.
33. Yakar L. ve Yavuz S. (2014). *Soru Kitapçıklarına Göre Değişen Madde Fonksiyonları.* IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Ankara. 9 - 13 Haziran 2014.
34. Yergin F. (2007). *Çoktan seçmeli testlerde madde sıralamanın madde ve test istatistikleri üzerine etkisi.* Yayımlanmış Yüksek Lisans Tezi. Mersin Üniversitesi. SBE.
35. Yılmaz H. B. (2014). *Madde ve seçenek sıralamasının değişen madde fonksiyonuna (DMF) etkileri.* IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi Ankara. 9 - 13 Haziran 2014.
36. Zenisky A. L. Hambleton R. K. & Robin F. (2003). *DIF Detection and Interpretation in Large Scale Science Assessments: Informing Item Writing Practices.* Center for Educational Assessment MCAS Validity Report No. 1. (CEA-429).
37. Zumbo B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa. ON: Directorate of Human Resources Research and Evaluation. Department of National Defense.