JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel

Hossein Karami*
(corresponding author)
Assistant Professor, University of Tehran,
Tehran, Iran.
Email: hkarami@ut.ac.ir

Ali Khodi**
PhD, University of Tehran, Kish international Campus,
Kish, Iran
Email: Ali.khodi@ut.ac.ir

ABSTRACT

Differential item functioning (DIF) is considered to be one of the tools for the examination of test fairness. This method is capable of finding the factors affecting the subjects' performance and prevent the occurrence of bias in the test. A plethora of methods for detecting Differential Item Functioning (DIF) has been suggested during the last couple of decades. The multiplicity of methods for diagnosing DIF might be a confusing issue for applied researchers and might lead to complications in the comparability of the findings of various DIF studies which have utilized different DIF detection techniques. This study aimed to investigate the comparability of results from three widely used DIF detection techniques: the Rasch model, Logistic Regression, and Mantel-Haenszel (MH). The data comes from an administration of the University of Tehran English Proficiency Test (UTEPT) which is a high-stakes test administered annually to PhD candidates. DIF analysis through the three techniques indicated that the three methods did not have significant differences in their performance. The Mantel-Hansel model flagged two items having DIF just similar to the findings of logistic regression model. Likewise, the items that were detected as strong-DIF items in Rasch model were the same as items detected by the two aforementioned models. Therefore, it might be concluded that use of different DIF detection techniques does not necessarily lead to flagging different items.

ARTICLE INFO

Article history:
Received:
9th, December, 2020

Accepted:
2nd, January, 2021

Available online:
Winter 2021

Keywords:

Differential Item Functioning, fairness, bias, validity

DOI: 10.22059/jflr.2021.315079.783

Khodi, Ali, Karami, Hossein (2021). Differential Item Functioning and Test Performance: a Comparison Between the Rasch Model, Logistic Regression and Mantel-Haenszel. *Journal of Foreign Language Research*, 10 (4), 842-853.
DOI: 10.22059/jflr.2021.315079.783

* Dr. Karami is an assistant professor of Applied Linguistics. His main research interests include various aspects of language testing and assessment.

** Ali Khodi is a PhD holder of applied linguistics who is in interested in language testing.

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 842 to 853

1. Introduction

It has become a truism to say that validity is the single most important consideration in test development and use (Bachman, 1990; Chapelle, 2016). Due to such significance, it is incumbent upon the test developers and users alike to ensure that their tests are valid. One of the threats against validity is the existence of construct-irrelevant variance (Messick, 1989). That is, performance on the test should not be affected by factors other the construct of focus in the test. Otherwise, the test would be biased (Karami, 2013a). A statistical technique widely applied to detect bias in the test is Differential Item Functioning (DIF).

DIF occurs when examinees with the same level of ability but from two different groups have different probabilities of endorsing an item. However, it is not synonymous with bias; rather, it is a prerequisite for bias. A biased item will certainly show DIF. On the other hand, an item which displays DIF is not necessarily biased. Thus, DIF is an essential but not sufficient condition for bias.

Various DIF detection techniques have been proposed in the literature (e.g. Clauser & Mazor, 1998; Kamata & Vaughn, 2004; Karami, 2012). These techniques do not always function similar in flagging items as DIF. So, it is essential to know how the results may be compared for studies using different DIF detection methods. The present study was an attempt to investigate the performance of three widely applied DIF detection techniques: the Rasch model, Logistic Regression, and Mantel-Haenszel.

2. Literature Review

Assessing psychological and cognitive properties of learners in education is usually done through administration of tests. The aim of these tests is to assess individuals in terms of the intended ability known as construct of measurement (Acar & Kelecioglu, 2010). The specification of individuals' abilities must be done through using a qualified instrument of measurement (i.e. exams, tests), and the contribution of any construct-irrelevant factor may reduce the utility of the test. Thus, tests in general and test items in particular should be able to measure ability without being affected by characteristics of subgroups that individuals belong to them (Uyar, Kelecioğlu & Doğan, 2017).

This happens because individuals' with equal abilities but from different groups should be able to answer the same items with similar probabilities. If in a given test there would be items that are probable to be endorsed by individuals of particular group more than other participants, the possibility of existence of bias is high (Cameron, et al., 2014). Although differential item functioning does not mean necessarily the existence of bias, it is an insufficient necessary for that (McNamara, & Roever, 2006).

The assessment of bias in the function of items could be done through examination of item responses for people with the same ability but different groups (e.g. age, race, gender, etc.). Therefore, in this examination the probability of the correct answers given is defined as differential item functioning (DIF) of an item (Steinberg, & Thissen, 2006). DIF is defined as the conditional probability of achieving correct answers of dichotomously scored items form people with the identical abilities but different demographical features.

There are two categorizes for DIF known as uniform and non-uniform. It is traditionally examined by comparing item responses for two sets of examinees named focal and reference groups. Although there is not a distinct rule for naming groups, arbitrarily the group minorities is named focal and the group of possibility advantaged majorities is labeled reference group (Steinberg, & Thissen, 2006). Uniform DIF is defined as the probability of responding correctly to an item uniformly higher for all level of ability of one group rather than the other group (Zumbo, 2003). While, non-uniform DIF is, considered to be known as crossing DIF.

For example, in a study by Zhu & Aryadoust (2020), the effect of participants' mother tongue and its relationship with DIF was examined. The findings showed that the test questions were not influenced by the native language of the learners. In another study, the effect of DIF was measured between German and English participants of the test of depression (Fischer et al., 2016). Findings indicated that despite the existence of this effect in 4 questions, its amount can be ignored.

By examining the bank of questions used for the entrance exam of the institute, Elena Oliveri et al. (2018) examined the existence of DIF. To that purpose, the effects of language level and ability, age, nationality, level and social class were examined. They used two methods of Mantel Hansel and IRT for the analyses. Their findings suggested that DIF happened mainly as a result of students' nationality. In another study by Chen, Liu and Zumbo (2019), a new method of differential action measurement has been proposed, which is based on the total score of the test and, of course, requires further measurement and evaluation. In addition, in this study the effect of each of these models on validity of the test results interpretation is addressed.

3. METHOD

Participants

The participants were 3000 applicants (both male and female) selected from the pool of examinees who had taken the University of Tehran's English Language Proficiency Test (UTEPT). The participants in the present were students of University of Tehran whose age ranged from 25 to 40. They were divided into two

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 842 to 853

groups based on their academic background: Humanities, and Science and Technology. Each group included 1500 examinees. An equal number of participants was selected for both groups so that sample size would not affect the results. Unfortunately, we did not have access either to the gender or age of the participants.

Instrumentation

The applicants to the Ph.D. courses of the University of Tehran are required to provide the authorities with a score in a proficiency test called the University of Tehran English Proficiency Test (UTEPT). The aim of the UTEPT is to identify those individuals who have the right level of English proficiency. The test is composed of three sections including Grammar, Reading and Vocabulary. All questions are in multiple choice format. The Reading section comprises passages immediately followed by a number of comprehension questions. The number of comprehension questions is different for each passage. Usually, a total raw score is reported to the candidates, which is simply the sum of scores they get on the three subtests. In this study, we only analyzed the Grammar part. There were a total of 35 items in this section.

Data Analysis

A variety of DIF detection techniques have been offered during the last three decades ranging from simple procedures based on difficulty indices (e.g. transformed item difficulty index (TID) or delta plot) to complex techniques based on Item Response Theory (IRT). Due to their conceptual elegance, IRT-based approaches are among the most widely applied DIF detection procedures. In this study three approaches were used for data analysis: the Rasch model, logistic regression and Mantel-Haenszel (MH).

In logistic regression, the item response is set as the dependent variable that should be predicted from other variables. The variables of the interest are the total score, the grouping variable, and the interaction between these two. In order to evaluate the model, three regression models can be determined where the first model includes only the total score, the second model both the total score and the grouping factor, and the last model these two factors plus the interaction term. The difference between the first and the third models should be first evaluated through the chi square test. If the chi square is significant, it shows the existence of either uniform or non-uniform DIF. The next step is to test the difference between the first and the second models for uniform DIF, and then compare the second and third models for non-uniform DIF. Jodoin and Gierl (2001) recommend the following guidelines:

1. Negligible or A-level DIF: $R2 < 0.035$
2. Moderate or B-level DIF: $0.035 \leq R2 < 0.070$
3. Large or C-level DIF: $0.070 \leq R2$

Mantel-Haenszel (MH) is a nonparametric DIF detection approach which rests on the idea of odds ratio. The odds ratio is obtained by

"pooling information across levels of the matching variable (typically observed total test score) to evaluate how much greater the likelihood of success is on a particular item for the reference group when compared to the examinees in the focal group" (Sireci & Rios, 2013, p. 175).

From among the extant IRT models, the Rasch model has gained a unique status due to its firm theoretical underpinnings and also its relation to conjoint measurement theory (for excellent expositions of conjoint measurement see Michell, 1990, 2003). Like other IRT models, the Rasch model focuses on the probability of endorsing item i by personm. Unlike other IRT models, however, the Rasch model essentially takes into account the person ability and item difficulty and considers item discrimination to be one. The probability of endorsing an item is modeled to be a function of the difference between person ability and item difficulty. The Rasch model provides us with sample independent item difficulty indices. DIF occurs when invariance is not accrued in a particular application of the model (Engelhard, 2008). That is, the indices are dependent on the sample who takes the test.

4. Results

With regards to the analyses conducted using logistic regression methods, Mantel Hansel and item response theory the indicators and indices of DIF are calculated and presented in the proceeding tables. Since the use of differential action models based on classical test theory has no specific assumptions, first in this section we explain the appropriateness of the data and assumptions of the models; next we compare models in terms of their capacities.

The first column in Table 1 shows item of the test. The second column (Measure) indicates the difficulty of each item (Lincare, 2010 b). For example, in this test, the most difficult item was number 9 and the simplest item was number 3. Regarding the third and fourth columns, it should be noted that primitively we check MNSQ to be between 1.3+ and -1.3, otherwise the next column, ZSTD, which refers to the scoring square. (Z-square) should be checked. If this index falls between -2 to2 the item has a good fit if not that item is considered to be over-fit (Linker, Lincare, 2010a) The same rules apply to columns 5 and 6. Likewise, columns 7 and 8 deal with point-measure correlations, which indicate a person's performance in answering and item as well as that person's overall ability. In the Expected column, the expected correlation after fitting the data is shown in the model.

Table 1. Model-Data Fit and assumptions

Item	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC Observed	PMC Expected	DIF Contrast
20	0.63	1.27	9.9	1.36	9.9	0.14	0.40	-0.32
10	0.17	1.19	9.9	1.25	9.9	0.21	0.39	-0.77
22	0.83	1.17	9.9	1.24	9.9	0.23	0.40	-0.32
34	0.42	1.18	9.9	1.24	9.9	0.22	0.40	-0.20
32	0.48	1.12	8.1	1.17	8.0	0.28	0.40	-0.33
1	0.45	1.11	7.7	1.15	6.7	0.29	0.40	-0.28
25	1.77	1.01	0.2	1.13	3.3	0.35	0.37	0.02
11	0.79	1.09	5.6	1.12	5.3	0.31	0.40	-0.13
18	-0.97	1.01	0.5	1.11	2.6	0.31	0.34	-0.29
14	-0.79	1.03	1.06	1.10	2.5	0.31	0.35	-0.04
31	-0.65	1.02	1.01	1.09	2.7	0.32	0.36	0.00
12	-0.89	1.02	0.9	1.08	2.0	0.31	0.34	0.05
35	-0.44	0.98	-0.9	1.04	1.4	0.38	0.37	0.09
27	1.37	0.97	-1.4	1.04	1.2	0.40	0.39	0.06
3	-0.02	1.02	1.4	1.02	0.9	0.37	0.39	0.00
5	0.52	1.01	0.6	1.01	0.3	0.39	0.40	-0.34
33	0.53	1.00	0.3	1.00	0.1	0.40	0.40	-0.06
15	0.33	1.00	-0.3	0.99	-0.5	0.40	0.39	0.04
13	-1.15	0.99	-0.2	0.95	-1.1	0.33	0.32	0.00
6	-0.20	0.99	-0.5	0.97	-1.0	0.39	0.38	0.10
9	1.66	0.89	-5.0	0.96	-1.1	0.46	0.38	0.00
29	0.25	0.95	-3.8	0.91	-4.2	0.45	0.39	-0.07
23	0.32	0.95	-3.9	0.94	-2.9	0.44	0.39	0.16
24	0.08	0.95	-3.8	0.93	-2.9	0.44	0.39	0.11
7	0.19	0.94	-4.2	0.93	-3.1	0.45	0.39	0.17
30	0.26	0.94	-4.6	0.91	-4.3	0.45	0.39	0.13
19	-0.67	0.94	-3.2	0.92	-2.4	0.42	0.35	0.23
21	-1.11	0.92	-3.2	0.85	-3.4	0.41	0.33	0.22
16	-0.32	0.91	-5.6	0.85	-5.4	0.46	0.37	0.35
4	-0.28	0.91	-6.0	0.84	-6.0	0.47	.038	0.29
26	-1.50	0.89	-3.7	0.76	-4.6	0.42	0.30	0.55
17	-0.35	0.88	-7.6	0.81	-7.1	0.49	0.37	0.00
2	-0.61	0.87	-6.8	0.78	-7.0	0.49	0.36	0.67
8	-0.22	0.86	-9.0	0.81	-7.7	0.51	0.38	0.18
28	-0.86	0.86	-6.5	0.76	-6.7	0.48	0.34	0.21

In the next step, to compare the power of each model in detecting differential item functioning in this analysis, questions 2 and 10 were identified as having intermediate level differential they were run and analyzed. The Mantel-Hansel model detected two items with DIF (based on chi-squares reported in Table 2). In a similar analysis using logistic regression based on the hypothesis of unidimensionality of the data the participants in the experimental and reference groups were matched and similar items were flagged as DIF. Finally, items 2 and 10 were selected by analyzing the data through item response theory. Despite the significant probability for some questions such as 5, 16 and 26, because the size of differential action was not significant, they were shown and categorized as questions with a negligible level of DIF.

It can be concluded that despite the use of different statistical methods, all three methods in the present study showed similar patterns in detecting DIF. What is clear is that the power of all three methods has been similar and largely the same from a diagnostic point of view.

Table 2. Comparison of Models in Detection of DIF

Item	IRT			Logistic Regression					Mantel-Hansel			
	ETS	Welch Prob.	DIF size	J & G	3rd R^2	2nd R^2	1st R^2	DIF χ^2	ETS	MH CHI	MH LOR	LOR SE
1	A	.0004	-.13	A	.109	.109	.107	4.066	A	0.0783	0.1347	2.8517
2	B	.0000	.56	B	.374	.374	.362	33.334	B	0.0964	0.5655	34.4344
3	A	1.000	.06	A	.185	.184	.184	3.622	A	0.0822	0.0461	0.2697
4	A	.0005	.20	A	.320	.320	.318	5.208	A	0.0892	0.1928	4.479
5	A	.0000	-.35	A	.209	.208	.201	22.725	A	0.0831	0.34	16.4146
6	A	.2044	.10	A	.214	.213	.213	2.759	A	0.0845	0.1076	1.5156
7	A	.0380	.09	A	.277	.276	.276	3.113**	A	0.0844	0.0969	1.2224
8	A	.0281	-.01	A	.377	.377	.377	0.4	A	0.0921	0.0005	0.0016
9	A	1.000	-.01	A	.301	.289	.289	30.642	A	0.1011	0.0136	0.007
10	B	.0000	-.51	B	.072	.072	.053	44.17	B	0.0789	0.5115	41.8188
11	A	.1201	-.01	A	.126	.124	.124	5.792	A	0.0803	0.0012	0.0007
12	A	.5703	.10	A	.153	.153	.153	1.268	A	0.0924	0.0994	1.0662
13	A	1.000	.01	A	.180	.180	.180	0.38	A	0.0998	0.0042	0.0001

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 842 to 853

Item	IRT			Logistic Regression					Mantel-Hansel			
	ETS	Welch Prob.	DIF size	J & G	3rd R^2	2nd R^2	1st R^2	DIF χ^2	ETS	MH CHI	MH LOR	LOR SE
14	A	.8059	.03	A	.145	.145	.145	0.061	A	0.0906	0.019	0.0272
15	A	.6114	.03	A	.217	.217	.217	0.559	A	0.0822	0.0351	0.149
16	A	.0000	.28	A	.317	.316	.313	10.56	A	0.0901	0.2739	8.9477
17	A	1.000	-.16	A	.358	.357	.356	6.749	A	0.0926	0.1629	2.9412
18	A	.0017	-.28	A	.158	.156	.152	8.155	A	0.0962	0.2821	8.3212
19	A	.0092	.15	A	.265	.265	.264	2.496	A	0.0929	0.1509	2.4866
20	A	.0001	-.01	A	.018	.017	.017	2.159	A	0.0764	0.0015	0.0003
21	A	.0236	.09	A	.278	.277	.277	2.41	A	0.1037	0.0743	0.4396
22	A	.0001	.-10	A	.068	.067	.066	4.263	A	0.0783	0.111	1.909
23	A	.0431	.11	A	.269	.268	.267	3.733	A	0.084	0.1249	2.0907
24	A	.1764	.05	A	.270	.269	.269	3.105	A	0.085	0.0596	0.4331
25	A	.8066	.12	A	.167	.167	.166	1.982	A	0.0946	0.1105	1.2683
26	A	.0000	.38	A	.330	.330	.326	9.928	A	0.1196	0.3623	8.9023
27	A	.5130	.10	A	.213	.211	.211	5.806*	A	0.0901	0.0866	0.8372
28	A	.0207	-.04	A	.374	.374	.374	0.198	A	0.1026	0.0301	0.0586
29	A	.3735	-.16	A	.273	.272	.270	8.195	A	0.0846	0.1564	3.2795
30	A	.1043	.07	A	.280	.279	.279	1.877	A	0.0851	0.089	1.0021
31	A	1.000	.02	A	.162	.160	.160	5.986	A	0.089	0.0175	0.0229
32	A	.0000	-.22	A	.115	.113	.110	11.327	A	0.0791	0.2215	7.6018
33	A	.4512	-.07	A	.209	.208	.208	2.052	A	0.0818	0.0535	0.3746
34	A	.0109	.03	A	.059	.058	.058	1.182	A	0.0773	-0.0268	0.0948
35	A	.2632	.07	A	.211	.211	.210	0.961	A	0.0873	0.0767	0.6969

Finally, it can be concluded that because all DFI flagged items are similar in three methods we can be sure about the soundness of these methods. According to the findings in Table 2, it can be concluded that Mantel Hansel method was the best determiner of items which were very difficult or easy. However, considering all the questions, the Rasch model is in the first place and then the logistic regression model. Another

finding of this study is that the percentage of questions with differential action is equal in all three methods. The similarities and capabilities of Mantel Hansel and logistic regression can be found in the similarity of their basic statistical model (CTT) (Hambleton & Rogers, 1989).

The effect of the existence of DIF on fairness of the test can be measured from the perspective of existing differences in the performance of the participants on the test. If the performance of the participants is affected by an irrelevant construct that would be probable that fairness is under question. In the present study, due to the small number of questions with DIF (2 questions), it can be stated that the condition for having a fair and valid assessment is met. Only 5.7% of the questions can be influenced by irrelevant factors.

It is clear that in general, a very large part of the present test was no biased against the participants and the fairness of the test was good in more than 95% of the grammar questions. However, it should be kept in mind that the concept of test fairness is multidimensional and its presence or absence is affected by various factors that should be controlled .

In order to discuss validity and DIF, based on what has been obtained we can say that the grammar section is one of the parts of language tests that shows the least amount of DIF and has good level of validity (McNmara & Rover, 2006).

Regarding the effect of background and field of study, it should be considered that there was no significant for them in assessing grammar (Hill, Hale, 1988). Also, the reason for the similarity of the results of the Rasch model with the findings of other two models can be the simplicity of the applied Rasch model. As in the studies conducted by Estaji & Zhale (2020) suggested may be having 2-parameter IRT models would show different results.

5. Conclusion

The purpose of this study was to investigate and explain the power of methods measuring DIF (Rasch model, logistic regression, Mantel Hansel). Another goal presented in this study was the effect of these methods on the validity and accuracy of statistical analyses conducted. Based on what was presented in Table 2 by comparing DIF methods, in recognizing questions with medium or high differential action level (which of course did not exist here), the Mantel-Hansel model offers appropriate efficiency.

It should be noted that the logistics regression model is a more general model than the Mantel-Hansel model; in Mantel-Hansel calculations only a model DIF is found that exists at all levels of a variable. In terms of statistical analyses and analysis time, the logistic regression model is more complex and superior than the other two models. It should be noted that for very simple or very hard items, the Mantel-Hansel method is assumed to be efficient, but for tests with moderate hardness, its efficiency is reduced.

Findings of this study can suggest new

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 10, Number 4, Winter 2021, Page 842 to 853

approaches for test designers, test organizers and especially analysts of test results to use the differences in the use of various methods of measuring DIF. Other languages components such as reading comprehension, which have structural differences with grammar, should also be examined to determine the dependence of the results and findings on the degree of difficulty or ability of the participants as influential factors. Perhaps by applying more complex models of IT, the capability of these theories in comparison can be substantially changed and reduced. However, it can be said that one of the most important aspects that could lead to DIF (Pae, 2004), known as the field of study, was measured in this study.

In terms of the computerized procedure of analyses it is recommended that the analyses could be done using different statistical packages in order to b sure the findings are true demonstration of the power that stems from each model rather than the software that was applied. In addition the number of the items existed in the study is another concern for the generalizability of findings that should be evaluated in various levels.

References

Acar, T., & Kelecioglu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory and Practice*, 10(2), 639-649.

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of applied psychology*, *77*(5), 598.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Cameron, I. M., Scott, N. W., Adler, M., & Reid, I. C. (2014). A comparison of three methods of assessing differential item functioning (DIF) in the Hospital Anxiety Depression Scale: ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. *Quality of life research*, 23(10), 2883-2888.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). New York: American Council on Education & Praeger series on higher education.

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

Chapelle, C. A. (2020). Validity in language assessment. *The Routledge Handbook of Second Language Acquisition and Language Testing*, 11.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2(2), 155–163.

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and Psychological Measurement*, *80*(3), 476-498.

Clauser, E. B. & Mazor, M. K. (1998) Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*.17, 31-44.

Elena Oliveri, M., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education*, *31*(1), 1-16.

Fischer, H. F., Wahl, I., Nolte, S., Liegl, G., Brähler, E., Löwe, B., & Rose, M. (2017). Language-related differential item functioning between English and German PROMIS Depression items is negligible. *International Journal of Methods in Psychiatric Research*, *26*(4), e1530.

Hale, G. (1988). Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing*, 5(1), 49-61.

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

Karami, H. (2012) An introduction to Differential Item Functioning. *International Journal of Educational and Psychological Assessment, 11(2),* 59-76.

Karami, H. (2013) The quest for fairness in language testing. *Educational Research and Evaluation, 19*(2&3), 158-169.

Linacre, J. M. (2010a). *A User's Guide to WINSTEPS®*. Retrieved May 2, 2010 from http://www.winsteps.com/ .

Linacre, J. M. (2010b). *Winsteps®* (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

McNamara, T. & C. Roever (2006). *Language Testing: The Social Dimension*. Malden, MA & Oxford: Blackwell.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, 11(4), 402.

Uyar, Ş., Kelecioğlu, H., & Doğan, N. (2017). Comparing differential item functioning based on manifest groups and latent classes. *Kuram ve Uygulamada Egitim Bilimleri*, 17(6), 1977-2000.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language testing*, 20(2), 136-147.

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, 1-25.