English Language in Focus
ELIF
e-ISSN: 2622-3597. Available online at https://jurnal.umj.ac.id/index.php/ELIF
Peer-review under responsibility of Muhammadiyah University of Jakarta, Indonesia

Gender Differential Item Functioning (GDIF) Analysis in Iran's University Entrance Exam

Soodeh Bordbar
English Department. Iran University of Medical Science. Tehran, Iran
ut.sbordbar@yahoo.com

ABSTRACT

The significant aspect of validity defines what the test score actually and potentially represents, especially to the causes of invalidity concepts of fairness, bias, injustice, and inequity. The Differential Item Functioning (DIF) examines the test items to define test fairness and to examine the validity of educational tests. If gender plays a major role in the testing items, this will lead to bias. This research examines the validity of a test for high-stakes and discusses gender's role as a bias in different linguistic tests, to explore validity and DIF analytics. To get a DIF analysis, the Rasch model had been used as a university entry requirement for English language studies for five thousand people taking part, who'd been randomly selected from a group of examiners participating in the National University Entrance Exam for Foreign Languages (NUEEFL), i.e., English literature, Teaching, and Translation. The test results indicated that the test scores are not free of construct-irrelevant variance, and certain inaccurate items have been modified following the fit statistics guidelines. Overall, NUEEFL's fairness was not clarified. These findings had been some advantage to test designers, stakeholders, administrators, and teachers through that kind of psychometric test. Then it suggested the future administering criteria and bias-free tests and teaching materials.

Keywords: Gender Differential Item Functioning analysis (GDIF), Bias, Dimensionality, Fairness, Rasch Model
Citation APA Style: **Bordbar, S. (2020). Gender Differential Item Functioning (GDIF) Analysis in Iran's University Entrance Exam. *English Language in Focus (ELIF)*, *3*(1), 49–68. https://doi.org/10.24853/elif.3.1.49-68**

INTRODUCTION

The significance and consequences of testing for language teaching, learning, and assessment have been widely questioned. The primary issue for test development and testing is the validity of scoring perspectives and the use of standardized tests. In the center of the language assessment, test use is taken into account. As suggested by Bachman (1990, p.55), the objectives of particular tests are intended to serve are the single most important consideration in developing language tests and in interpreting their results. He also argues that tests are not developed and used in a value-free psychometric test tube; it is often intended to serve the requirements of the education system or the entire society (Bachman, 1990, p. 279).

Looked at from a different perspective, it is more important to search and find evidence when the stakes of the test increase. Understanding the significance of the consequences of the test can significantly differ among stakeholder groups; those who have to experience, the consequences are more serious than those who do not suffer (Kane, 2013, p. 48). High-stakes testing is one of the most provocative aspects of education, and the technicalities involved are highly complex. High-stakes and teacher-made tests vary in the results' interpretations and consequences. In the teacher-made test, scores and exams are interpreted in different ways. Failing in teacher-made tests could be interpreted as a failure to learn the materials, whereas passing the test indicates mastery of the subject.

In the Education Reform glossary (2014), the data gathered in high-stakes tests are used to administer punishments. Similarly, high-stakes tests strive to use tests and assessments on their own to take actions with major educational, financial or social consequences (Genesee & Upshur, 1996, p. 6). The aim is to give all participants the same opportunity to analyze the outcome of the tests in high-level testing (Song & He, 2015). In particular, the American Educational Research Association (AERA) released guidance on high-stakes tests for policymakers to enhance education and highlight cautious assessments to prevent substantial damage from the tests (Dunne, 2015).

It is also extremely important that the test is fair to different test participants' groups. In other words, the test should not be tempted by the characteristics of test-takers like gender, ethnicity, academia, etc. Gender is one of the factors commonly known as a source of construct-irrelevant variance. If the impact of gender becomes significant, there are increasing controversies about the preference of the tests. And this problem certainly diminishes the validity of the test. Moreover, a statistical method must examine such a problem to determine whether the test items differ between the test-takers and, at least, to investigate the source of construction invariance. Differential Item Functioning (DIF) by using the Rasch model measurement is one of the recommended methods for achieving this goal.

The DIF technique is a useful way to classify potentially problem entities (Angoff, 1993). Moreover, in second-language evaluations (L2), DIF analysis can be used to particular significance. DIF happens where the properties of an object vary from those in another group in one group (Furr & Bacharach, 2007, p. 331). For this reason, Furr and Bacharach (2007) use the DIF example when an object has various levels of difficulty for men and women. In the DIF processes, the test item works equally for two or more groups of test participants typically identified according to their race/ethnic background, gender, age/experience or disability (Scheuneman & Bleistein, 1989, pp. 255-256).

A range of potential methods are also available, but only a small number are currently being used. For a comprehensive explanation of DIF detection methods, see the following: Sheuneman and Bleistein, 1989; Wiberg, 2007. 2004; McNamara and Roever, 2006. As a German and applicable research

method, Rasch-measurement helps in designing and adjusting a measurement instrument and in calculating measurements for parameter statistical research (Boone, Staver, & Yale, 2014). It implies that psychometric analyzes are used. Moreover, unidimensionality, local independence, and model fit are the three presuppositions in the Rasch model.

In the National University Entrance Exam for Foreign Languages (NUEEFL, in Iran), a large number of test-takers also take seating each year. English is one of the subjects in the exam. To take various university degrees, the students must also take several high-stake tests. In Iran, high-stake tests are typically used too. Numerous studies have examined multiple perspectives of these tests (i.e., Farhady & Hedayati, 2009; Salehi & Yunus, 2012 a, b; Mirzaei, Hashemian, & Tanbakooei, 2012). For instance, Tahmasbi and Yamini (2012) investigated the teachers' perspectives on student scores and their effect on the prospects' lives of those who took the Iranian University Entrance Exam (IUEE). The findings indicate that high school teachers play no role in the development and maintenance of IUEE. As teachers pointed out, neither language skills nor knowledge was involved. They assumed that the use of test-taking skills, tactics, and strategies is primarily a rationale for effective or failed high-risk tests.

Further research by Sadeghi (2014) evaluated the influence of high-stakes testing on TOEFL and IELTS preparatory courses in Iran, using the Structuration Theory and Washback Hypothesis. He researched an interpretive ethnographic case by way of observations and field notes to find out how high-stakes tests influence teachers' curriculum and methodology. On closer inspection, the teachers constantly encountered challenging questions which led to variations in their response to exam pressures.

The purpose of this paper is to find due to gender performance among language test subtests. This is achieved by providing the validity of the proficiency test psychometrically. The present study addresses the following research questions:

1. Do the items of the test fit the Rasch model?
2. Is there any case of local item independence and unidimensionality among all the subtests of NUEEFL?
3. Is participants' gender a source of DIF in the subtests of the NUEEFL?

RESEARCH METHODOLOGY

The participants (N = 5000) were selected from among the pool of examinees from a population of 20,000 who had taken a recent version of the NUEEFL test. They were selected randomly from the two gender groups (i.e., males and females). Of the 5000 participants, 3335 were female, and the remaining 1665 male. The academic background and the age of the participants were not an issue.

The first instrument used in this research is the NUEEFL. It consists of a total of 95 items of which 25 are general English questions (from # 76 to 100). The other 70 items come under six subtests:

a. Grammar (10 items)-(from # 101 to 110)
b. Vocabulary (15 items)-(from # 111 to 125)

c. Sentence Structure (5 items) -(from #126 to 130)
d. Language Functions (10 items) -(from # 131 to 140)
e. Cloze Test (15 items) -(from #141 to 155)
f. Reading Comprehension (15 items) -(from # 156 to 170)

The NUEEFL test is annually administered to more than 100,000 applicants who want to get a bachelor's degree in foreign languages from a public university. All questions are in multiple-choice format and scored dichotomously. The test is time-restricted, lasting 105 minutes. In NUEEFL the correction for guessing is applied in the test items. In other words, guessing is not allowed with a total of three incorrect answers offsetting a correct answer.

Another instrument used for analyzing the data is Winsteps software (Version 3.92.1 updated in February 2016) (Linacre, 2016a, b). Winsteps constructs the Rasch measures using simple data sets (i.e., usually of persons and items) and applies the dichotomous Rasch model. This software can analyze combined item types, for instance dichotomous, multiple-choice, and multiple rating scales. It also examines in-depth the structure of the items and persons. It uses a powerful diagnosis of multidimensionality via principle components analysis of residuals to detect and quantify substructures in the data (Linacre, 2016a). The version of Winsteps software (Version 3.92.1) used here can process up to 9,999,999 persons, 60,000 items, and up to 255 categories. It should be noted that Winsteps does not consider missing data or non-administered items.

Winsteps implements two methods of estimating the Rasch parameters, a) Joint Maximum Likelihood Estimation (JMLE), b) PROX (the Normal Approximation Algorithm devised by Cohen in 1979). In keeping with the methodology of the present study, Winsteps software implements the JMLE method for estimating the Rasch parameters. In the JMLE formula, the estimate of the Rasch parameters happens when the observed raw score for the parameter matches the expected raw score. And the word "Joint" in JMLE means the simultaneous occurrence of the estimation of items and persons and rating scale structure of data matrix (adapted from Winsteps manual by Linacre, 2012; 2016a).

This study focuses on the validity issue assessed through the application of the Rasch model. Given the nature of the study, the statistical and mathematical assumptions must be met. The steps followed are outlined below:

1. Preparing the data file for analysis, using SPSS and *Winsteps* software
2. Checking the data-model fit
3. Checking the assumptions of the Rasch model including, dimensionality and local independence
4. Analyzing DIF in the whole test and across the subtests

FINDING AND DISCUSSION

Reliability analysis is a crucial part of any assessment. The *Winsteps* output provides the model requirements of an unbiased reliability estimate. The *Winsteps* tables provide a wide range of indices that can be used to evaluate the reliability of an instrument. The item

separation index is from 0 to infinity, and the reliability index is from 0 to 1. Reliability analysis is sensitive to the sample size which varies with the number of items and participants. As Linacre (2012) has said, the low item reliability means that "your sample is not big enough to precisely locate the items on the latent trait" (p. 644). The result of reliability analysis showed a reliability value of 1.00 which has high reliability for 95 test items. We can conclude with confidence that the high item reliability was also affected by the sample size of the data.

Checking the data-model fit estimate

The *Winsteps* software normally assesses the fit of the model through obtained statistics indicators of mean-square fit values (MNSQs) and the standardized Z values (ZSTDs). The values in the range of MNSQs are considered from zero to infinite (0- ∞) and the expected value is 1. Values above 1 likely show a deviation from the unidimensionality, and values less than 1 indicate an overfit in the response patterns with the data-model. The overfit in the model implies the existence of dependency among responses or items. Values between 0.70—1.3 are considered acceptable or so called good fit values and values less than 0.70 indicate overfit. Meanwhile, values above 1.3 signify underfit. Overfit for a low ability group indicates that the item is more discriminating between slight differences in ability.

In analyzing the model fit estimation, it is necessary to eliminate the participants with a total score of zero. The data were screened for outliers. As indicated earlier, the Rasch model does not estimate the zero scores and inevitably they are omitted from the analysis process. Hence, from a total of 5000 participants, 4965 remained for data analysis. Besides, the fit indices should be reported for the item calibration. The estimation of difficulty parameter and model fit estimations revealed that the range value of difficulty parameter is from 2.45 to -2.99, with a mean score of 0, and Standard Deviation (*SD*) of 1.21. The most difficult item is item Q.155 (Measure = 2.45) and the least difficult is the item Q.87 (Measure = -2.99).

It appears that 26 items were not located in the acceptable range. The range value of the outfit-MNSQs varies from 0.57 to 3.3 which denote that these items do not fit the model. The investigation of item statistics of outfit-MNSQs reveals that 26 items equal to 27% of items (i.e., 155, 126, 137, 105, 118, 166, 101, 103, 158, 111, 115, 133, 121, 167, 109, 128, 122, 156, 99, 84, 153, 149, 108, 160, 91, and 79) out of a total 95 items were not located in acceptable rating scale of 0.70 to 1.3. There are four items from a total of 25 items in General Questions (i.e., 79, 84, 91, and 99), from total of 10 items, five items in Grammar (i.e., 101, 103, 105, 108, and 109), from total of 15 items, five items in Vocabulary (i.e., 111, 115, 118, 121, and 122), two items out of five items in Sentence Structure (i.e., 126, and 128), two items out of 10 items in Language Functions (i.e., 133, and 137), three out of 15 items in Cloze Test item (i.e., 149, 153, 155,), and from a total number of 15 items, five items in Reading Comprehension (i.e., 156, 158, 160, 166, and 167) did not fit the model. The presence of a large number of misfitting

items reveals that the data does not fit the model in each subtest. Therefore, the model and its assumptions may be violated. It is possible that the Rasch model unidimensional assumptions also may or may not attain the desirable results.

Checking unidimensionality and local independence

There are multiple methods for assessing dimensionality, including the data-model fit statistics. However, several studies have reported that these statistics do not have the ample sensitivity necessary for detecting multidimensionality. Besides checking the data-model fit, it makes sense to employ the Principal Components Analysis (PCA) on the raw data and residuals. In the present research, the Principal Components Analysis of residuals and a series of *t*-tests were used to check the unidimensionality of the test. The following criteria were used to determine the unidimensionality of the test through the PCA analysis: a) if the amount of variance explained by measures be > 60%, b) "the unexplained variance of the eigenvalue for the first contrast (size) < 3.0 and unexplained variance explained by first contrast < 5% is good" (Linacre, 1991-2006, p. 272). Also, with regards to the criterion for eigenvalue, the expected eigenvalue is less than 2.0, but, in practice, a secondary dimension in the data usually requires a value of 3.0 or more.

In order to hold unidimensional, there must be little residual correlation among items remaining. It is presumed that unidimensional can be supported if 5% of *t*-tests are significant. If the level of significance of *t*-test is more than 5%, the local independency and unidimensional will be violated.

The amount of the variance explained by different components in the data is 34.8% of which 12.7% is explained by persons and 22.1% by items. This indicated that a dominant first factor is present. As a rule of thumb, the variance explained by the first factor should be greater than 60% to be indicative of unidimensionality (Linacre, 2006). The result obtained here (34.8%, with an eigenvalue of 50.70) is lower than the minimum level necessary to demonstrate a unidimensional trait in the data. This showed that the items did not fit the model with item-person leveling. The first, second, third, fourth, and fifth unexplained variance with the eigenvalues of 3.4, 2.5, 2.2, 1.9, and 1.7 which were satisfactory according to the criteria. The results of the data analysis suggested that the unidimensionality does not hold across the whole test.

Local independence was examined through checking the ability parameter in order to identify whether the responses to items could be independent of each other (Pae, 2011). The Benjamini-Hochberg method was used to investigate the Rasch assumptions (Benjamini & Hochberg, 1995). First, the item difficulty parameter for all items was calculated using the Rasch model (it is called Level A). The investigation of outfit MNSQs statistics showed that 20 items had a value greater than 1.3. These 20 items confirm the unidimensional assumption in data analysis. Second, the item difficulty

parameters for these 20 items were determined using the Rasch model (it is called Level B).

The total sum of the differences between the difficulty parameter of these 20 items is -1.124 which is calculated from levels A and B. The constant correction value is -0.056 which is obtained by dividing -1.124 to the number of items (i.e., 20). As the next step, the ability parameter based on items in levels A and B was calculated. The results indicated that from a total of 5000 Student's *t*-statistics, 2680 or 53.6% were significant meaning that they were larger than the acceptable level of 5%. Thus, it is concluded that the unidimensionality and local independence assumptions do not hold in the entire test. To appreciate if unidimensionality and local independence hold in each subtest, the results of item calibration in each of six subtests separately were analyzed.

In PCA, there are multiple ways to determine the factorability of inter-correlation matrix and to assess the appropriateness of using exploratory factor analysis. To determine the number of factors, some considerations such as Kaiser's criterion, Cattell scree-plot, and total variance explained should be taken into account. As shown in Table 1., for instance in Grammar section the Kaiser-Meyer-Olkin measure of sampling adequacy was 0.909 which is above the recommended value of 0.7. Also, the Bartlett's test of sphericity was significant (p- value = 0.00, $p < .05$).

Table 1. KMO and Bartlett's test in Grammar

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.909
Bartlett's Test of Sphericity	Approx. Chi-Square	9976.549
	df	45
	Sig.	0.000

The results revealed that all the extracted factors are not of interest to the researcher. The purpose of factor analysis is to explain the components with the smaller number of factors from the primary variable. First, it seeks to determine the number of factors or components that are kept in the factor analysis. In order to keep factors, usually mathematical criteria such as, Kaiser's criterion or Cattell Scree-plot test are employed. The Kaiser's cut-off value specifies the number of factors which have an eigenvalue of 1 or higher. Only those factors are kept which have a sum of squared factor loading (eigenvalues) equal or greater than 1. Some researchers keep enough factors to explain 80% of the variation. As shown in Table 2., the only factor with the value of 3.719 was component 1.

Table 2. Total variance explained in Grammar

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.719	37.193	37.193	3.719	37.193	37.193
2	.898	8.979	46.171			
3	.765	7.648	53.819			
4	.724	7.237	61.056			
5	.717	7.168	68.224			
6	.694	6.943	75.166			
7	.661	6.615	81.781			
8	.649	6.494	88.276			

9	.601	6.012	94.288
10	.571	5.712	100.000

The Cattell scree-plot displays the eigenvalues associated with a component or factor in a simple line plot in a diminishing size pattern. This scree-plot can be employed to graphically assess the optimal number of factors and to visualize factors which show most variability in the data. The ideal configuration in scree plot is a steep curve, followed by a bend and then a flat horizontal line. The place where the curve pattern for eigenvalues goes horizontal is called the scree point. The screen test shows the place where the smooth decrease of eigenvalues appears to level off to the right part of the plot. In the right side, the factorial scree is found. Put simply, the real operating factors are located on the left and the error operating factors are on the right (See Figure 1) (Ledesma et al., 2015; Raîche et al., 2012).

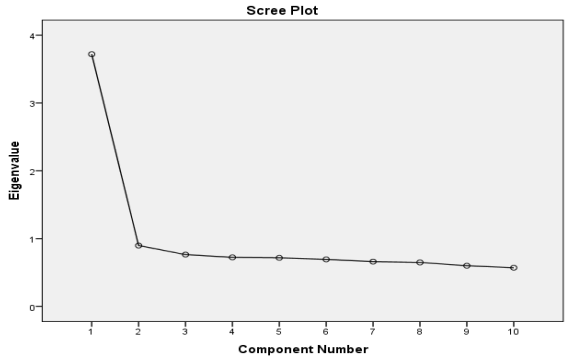

Figure 1. The scree-plot in Grammar

Regarding this, after selecting the components, we should perform factor matrix rotation. The main purpose of the rotation is to make the output more understandable by finding a simple structure. Rotations can be orthogonal (i.e., independence) and oblique (i.e., dependence/related). In a nutshell, the rotation was orthogonal and the results of component factor analysis displayed only one extracted component in a way that all variables were related to the first factor (See Table 3).

Table 3. Component matrix in Grammar subtest

	Component 1
X101	.534
X102	.640
X103	.648
X104	.600
X105	.643
X106	.554
X107	.588
X108	.652
X109	.608
X110	.620

Further, in each question of all subtests, extracting of the main factors and the PCA analysis were run. As a rule of thumb, a PCA analysis is significant if required variable for explaining 70% of variance is less than half of the variables. In the Grammar subtest the variance explained by the first factor was 37.193% is lower than the requirement of this criterion. It does not determine a unidimensional trait.

It can be inferred from the results of the PCA analysis in each subtest that the exploratory factor analysis was not significant for all six subtests. All in all, the results of the PCA analysis in the entire test were not significant. Besides, the unidimensionality of the trait in the data did not confirm. Moreover, there was no solid evidence of local independence across subtests and in entire test.

DIF analysis

Test developers use several quality controls or statistical procedures

to make sure that the test items are appropriate for and fair to all examinees (Camilli & Penfield, 1997; Holland & Wainer, 2012; Ramsey, 1993). In this study, DIF analysis among all items across the entire test and subtests of the NUEFFL between male and female participants were investigated. When analyzing DIF in the Rasch model, it is necessary to examine both the magnitude of the difference in logit units between groups and the statistical significance of the difference (Linacre, 2016a).

The magnitude of the DIF value should be at least 0.5 logits. In this phase of the study, DIF analysis was tested between groups of males and females. In order to examine the invariance, it is imperative to inspect the difference between the DIF analyses of these two groups by gauging the *t*-tests of the statistical significance of the data. For statistically significant DIF, the probability of such differences (0.5 logits or larger) exists at random meaning ≤ 0.05. It is probable that such differences might crop up in the absence of systematic item bias among the test items (Linacre, 2016a).

The results obtained here revealed different difficulty levels in 85 items. They also showed that in 41 items the DIF contrasts were negative which means that these items are more difficult for female participants, whilst in 10 items the DIF contrast was zero. Further, in 44 items DIF contrasts were positive, demonstrating that these items were more difficult for the male group. The item difficulty has a normal distribution between gender groups.

Moreover, in 40 items the statistical differences between compared groups were significant. DIF Measure reports the difficulty (diff.) of each item for each person classification. In other words, DIF Measure is equal to DIF size plus the overall item difficulty (Linacre, 2016b). The difference in difficulty level showed that a large number of items were located above zero (See Figure 2). It denotes that the difficulty level of items in the NUEEFL was large. As proved by the DIF results, the NUEEFL test was a difficult test for the participants. And, the invariability of questions in gender groups was not accepted. Thus, the statement that the participants' gender is not a source of DIF in the NUEEFL is rejected.

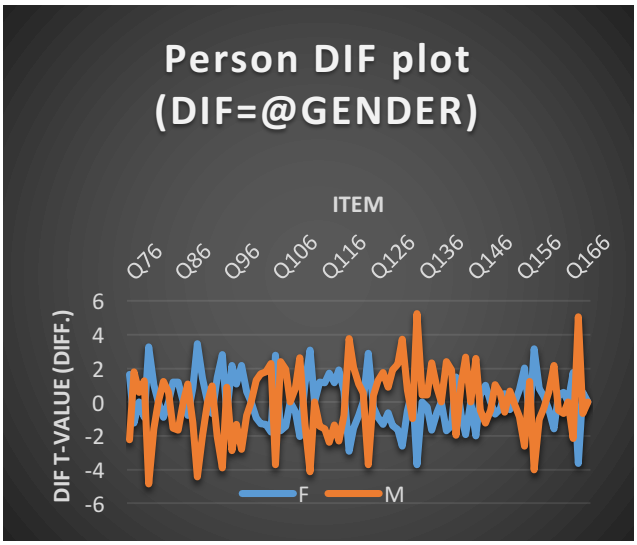

Figure 2. Comparing t-value difference between groups of males and females

However, an item that displays DIF is not essentially unfair to various groups of participants. And, it is too hasty to claim that the NUEEFL test is unfair to both male and female participants. In sum, finding the significance DIF between gender groups, the NUEEFL appeared not to be a DIF-free person estimates test. Hence, it is concluded that there is difference between males and females in answering the NUEEFL test. The following

paragraphs present the results of the DIF analysis for the subtests. For instance, in the Reading Comprehension subtest the DIF analysis showed that 5 items out of 15 have significant DIF. As shown in Table 4, the highest variability of DIF contrast in the Reading Comprehension subtest is related to the item 168.

Table 4. DIF results for Reading Comprehension items

Entry Number	Item	Sub Group	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch t	Rasch-Welch df	Rasch-Welch Prob
81	Q156	RC	M	0.44	F	0.54	-0.1	-1.12	INF	0.2624
82	Q157	RC	M	-0.04	F	0.24	-0.28	-3.3	INF	0.0010
83	Q158	RC	M	1.56	F	1.37	0.19	1.55	INF	0.1206
84	Q159	RC	M	0.47	F	0.97	-0.5	-5.12	INF	0.0000
85	Q160	RC	M	-0.28	F	-0.17	-0.1	-1.29	INF	0.1987
86	Q161	RC	M	0.62	F	0.66	-0.05	-0.5	INF	0.6143
87	Q162	RC	M	1.07	F	0.99	0.08	0.76	INF	0.4497
88	Q163	RC	M	0.05	F	-0.18	0.22	2.69	INF	0.0072
89	Q164	RC	M	0.2	F	0.26	-0.06	-0.66	INF	0.5119
90	Q165	RC	M	0.2	F	0.28	-0.08	-0.88	INF	0.3784
91	Q166	RC	M	1.57	F	1.57	0	0	INF	1.0000
92	Q167	RC	M	0.96	F	1.27	-0.3	-2.8	INF	0.0052
93	Q168	RC	M	-0.42	F	-0.89	0.47	6.22	INF	0.0000
94	Q169	RC	M	1.23	F	1.34	-0.1	-0.88	INF	0.3779
95	Q170	RC	M	1	F	1	0	0	INF	1.0000

Note1. Highlighted are the DIF flagged items in Reading Comprehension subtest.

Note2. RC = Reading Comprehension; M = Male; F = Female; INF = Infinity

Figure 3 displays the variance of ordering and spacing of all items. Except for two items (e.g., item 157 & 163), the remaining items have the same direction in ordering between compared groups of males and females. The results show that item 163 was an easy question in favor of the female group whereas item 157 is an easy question for male participants; it is interpreted as a male-favoring item. Except for items 157 and 163, the rest of DIF-Flagged items have the same ordering direction. And the difference of variability among items has been considered with the difficulty level of items in each subtest.

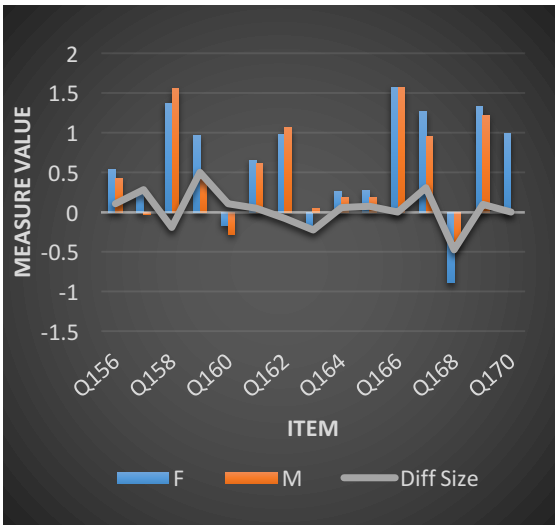

Figure 3. DIF size for Reading Comprehension items across gender groups

The results revealed that among 5 DIF-flagged items in the present subtest, the female group signified a lower degree of difficulty in 3 items. And, males indicated a lower degree of difficulty in

the 2 other items. Thus, it is possible to conclude that the questions in the present subtest are easier for the female participants in comparison with the other group.

As discussed earlier, DIF analyses deliver a partial answer to fairness issues. Therefore, if grouping happens in a test and the items are favoring one group, then the test may not be fair enough for the other group. Hence, DIF analyses were investigated in the NUEEFL to resolve the matter. Among all subtests, General Questions, Sentence Structure, Language Functions, Cloze Test, and Reading Comprehension turned out to be more female-favoring, whereas Vocabulary and Grammar were more favored by male participants (Please see Appendix for the DIF size figures in other Subtests). Our investigation has shown that a correct answer may require other knowledge, ability, and skill than the ones that the items aim to measure. Furthermore, the data analysis showed that these additional skills or knowledge are not equally presented between groups of males and females. And the items did not lead to the construct validity of the test for all the participants in gender groups because in some subtests the values obtained through DIF analysis were greater than expected (Uiterwijik & Vallen, 2005).

It has to be emphasized that four items (i.e., 108, 142, 157, & 163) were controversial in terms of the variance of direction in ordering. It can be inferred that these four items can be the source of bias. Moreover, an unequivocal result is that the favoring happens in the NUEEFL items and among subtests. Thus, the test is partial for the disfavored group.

CONCLUSION

There is an ongoing interest in comparing cultural, ethnic, or gender groups. DIF studies are absolutely essential in high- stakes testing programs. Furthermore, possible gender and/or ethnicity bias could negatively impact one or more groups as an irrelevant construct. In fact, the test administers attempt to develop perfectly fair test batteries; however, the dearth of research on the NUEEFL test makes it impossible to assess its fairness accurately.

The NUEEFL taken by tens of thousands of students annually acts as a gate-keeping test for those aspiring to enter the higher education system in Iran. In line with the main purpose of this research, DIF analysis was used to identify bias items across gender groups. It did not confirm a similar probability of endorsing the test items. The results of the study indicated that 40 items out of 95 turned out to be DIF-flagged items. This suggests that NUEEFL test scores are not free of construct-irrelevant variance. Hence, it does not support the argument for the construct validity.

To the best of the researcher's knowledge, the investigation of DIF analysis on the National Organization for Educational Testing's (NOET) data, the NUEEFL test, across gender groups was a brand new research project. This was particularly the case because the DIF analysis used the Rasch model in all sections of the national test. Whilst there are studies which investigate DIF across gender groups in language tests administered in Iran, they are principally concerned with the University of Tehran English Proficiency Test (UTEPT). Mohammad et al., 2014; Rezaee and

Shabani, 2010; Salehi and Tayebi, 2011 are examples of such endeavors.

Our findings are in line with parts of the research performed on the UTEPT. In Amirian et.al. (2014) and Rezaee and Shabani (2010) DIF were displayed and observed in the different sections of the UTEPT. Although in some studies the results obtained were not compatible with our findings. For instance, Salehi and Tayebi (2011) have not found DIF items in reading section of the UTEPT. In another study Rayan and Bachman (1992) examined DIF with respect to the participants' performance in the TOEFL and the FCE exams. The results showed that the difference in the performance of males and females at the item level was negligible, whereas Carlton and Harris (1992) in a gender focused DIF study found that the females performed better than males. The results of the present research were compatible with Carlton and Harris's.

While that there are many gender focused DIF studies in a first language setting (i.e., Li & Suen, 2013; Ryan & Bachman, 1992; Zhang, Dorans, & Matthews-Lopez, 2003), few deal with DIF analysis in the context of English as a Foreign/Second Language (EFL/ESL) e.g., Alavi et al., 2011; Pae, 2004; Rezaee and Shabani, 2010; Salehi and Tayebi, 2012.

Rezaee and Shabani (2010) found significant DIF between gender groups in the UTEPT. The result of their study was verified by Karami (2011) who used the Rasch model to examine gender DIF in the UTEPT. The result of the study revealed that only 3 among 19 DIF items displayed practical significant DIF. Also, Amirian et al., (2014) detected gender DIF with the UTEPT. They performed a twofold research. The result of the first phase revealed that there is substantial DIF between gender groups in UTEPT. In the second phase, they performed content analysis on the DIF-flagged items to understand the source of DIF; and the result showed that humanities-oriented topics were mainly female favoring, while science-oriented topics were mostly favored by males. In fact, the literature on gender DIF in EFL/ ESL context using the Rasch model is inadequate (Karami, 2010, 2015).

Moreover, several studies have investigated DIF at the language skills level. For instance, Aryadoust, Goh, and Kim (2011) examined gender DIF in the Michigan English Language Assessment Battery (MELAB) listening section using the Rasch measurement. The result of the uniform DIF (UDIF) revealed that two items favor different gender groups. Also, the non-uniform DIF (NUDIF) analysis showed several items with significant DIF mostly favoring low proficient male participants.

To bridge this gap, the present research attempted to validate the NUEEFL test in the case of gender. The present gender DIF study implements the Rasch model to figure out whether the NUEEFL as a high-stakes test shows substantial DIF in favor of a specific gender group. The results indicated that the NUEEFL test was favored by females. Apparently, disfavored group (i.e., male group) was not treated fairly. The test turned out to be unfair.

The results of the present study are controversial due to the statement of Rezai-Rashti and Moghadam (2011). They believed that a range of restrictions have been put on the number of female

students that can enroll in each field to restrict females altogether from certain majors; whereas the result of the present research confirmed that the test is administered in favor of females.

With respect to research conducted by foreign scholars, the work of Lin and Wu (2003) is very similar. They employed the computer program CIPTEST for DIF/DBF (Differential Bundle Functioning) analysis and examined dimensionality of the English Proficiency Test in China. Their work shows much greater gender DIF in the overall test and among subtests. In another study, Tae (2004) conducted DIF analysis in the Reading Comprehension section of Korean National Entrance Exam for Colleges and Universities whose results showed gender DIF in Reading Comprehension section of the exam.

Additionally, the results of the present study are consistent with Karami's (2015) with respect to dimensionality of NUEEFL, in which the multidimensionality in the whole test and among sub-tests was proven to exist. The results of the present research showed that the NUEEFL is not unidimensional.

Fairness and DIF analysis are broad concepts which involve analyzing the DIF among the items, considering the performance of different groups, or focusing the bias on the performance of every individual participant (See Camilli, 2006; Xi, 2010; Kunnan, 2010). Test bias is related to the issue of test fairness, and the social ramifications of test results have unfairly advantaged or disadvantaged specific groups of test takers.

The university entrance exam often raises concerns about the issues of test fairness and bias. Since Messick (2013) has noted that validity and fairness are a matter of degree. As Messick (2013) has pointed out, a test is neither absolutely valid nor absolutely invalid. Also, the fairness of the test is not absolute. A test can be to some degree fair or unfair. The results of the present study show that the DIF were significant among 40 items. Among subtests of NUEEFL (i.e., Sentence Structure, Language Functions, Cloze Test, and Reading Comprehension) turned out to be more female-favoring. Whereas Vocabulary and Grammar were more favored by male. And among DIF items, four items (i.e., 108, 142, 157, & 163) had critical results regarding the variance of direction in ordering. And the construct validity of the NUEEFL was threatened in compared groups.

It should be noted that in present research only receptive skills including reading comprehension, vocabulary, and grammar, were examined. Other skills such as listening, writing, and speaking were not included. In listening comprehension, for instance, most females were found to have an advantage compared to males (Boyle, 1987; Cole, 1997). The results of analysis indicated significant DIF between males and females and the findings from this study are not consistent with the results of Ryne and Bachman's (1992) who found no gender difference in any of TOEFL subtests.

REFERENCES

Alavi, Mohammad, AliRezaee, Abbas, & Amirian, MohammadReza. (2011). Academic Discipline DIF in an English Language Proficiency Test. *Journal of English Language*

Teaching and Learning, *7*(5), 39–66. http://noo.rs/KIrXf

Aryadoust, V., Goh, C. C. M. & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8(4),* 361–385.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing . *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). *Rasch Analysis in the Human Sciences*. Springer Science, and Business Media. https://doi.org/10.1007/978-94-007-6857-4

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning, 37*(2), 273-284.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., Vol. 4, pp. 221-256). Westport, CT: American Council on Education & Praeger.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on the Mantel-Haenszel log-odds ratio. *Journal of Educational Measurement, 34,* 123–139.

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons.* Princeton, NJ: Educational Testing Service.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *The British Journal of Mathematical and Statistical Psychology, 32,* 113-120.

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service.

Dunne, D. W. (2015). *Cautions Issued About High-Stakes Tests | Education World*. Education World. https://www.educationworld.com/a_issues/issues110.shtml

Furr, M. R., & Bacharach, V. R. (2007). *Psychometrics: An Introduction*. Thousand Oaks, CA: SAGE.

Holland, P. W., & Wainer, H. E. (2012). *Differential item functioning*. London, UK: Routledge.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Karami, H. (2010). A differential item functioning analysis of a language proficiency test: an investigation of background knowledge bias. Unpublished MA Thesis, University of Tehran.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5(2),* 27-38.

Karami, H. (2015). A closer look at the validity of the University Entrance Exam: Dimensionality and generalizability. (Unpublished Ph.D dissertation, University of Tehran).

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183–189.

Ledesma, R. D., Valero-Mora, P., & Macbeth, G. (2015). The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *The Spanish Journal of Psychology*, *18*, E11. https://doi.org/10.1017/sjp.2015.13

Li, H., & Suen, H. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing. 30,* 273-298. https://doi.org/10.1177/0265532212459031.

Lin, J., & Wu, F. (2003). Differential performance by gender in foreign language testing. *Paper presented at the annual meeting of the national council on measurement in education (Chicago, IL.).*

Linacre, J. M. (1991-2006). A user's guide to Winsteps® Ministep Rasch-model computer programs. Retrieved January, 10, 2007, from http://www.winsteps.com/aftp/winsteps.pdf

Linacre, J. M. (2006). Data variance explained by measures. *Rasch Measurement Transactions, 20,* 1045–1047.

Linacre, J. M. (2012). A user's guide to Winsteps [User's manual and software]. Retrieved from http://www.winsteps.com/winsteps.htm.

Linacre, J. M. (2016a). Winsteps® Rasch measurement computer program User's Guide. *Beaverton, Oregon:* Retrieved from http://www.winsteps.com/

Linacre, J. M. (2016b). Winsteps® (Version 3.92.1) [Computer Software]. *Beaverton, OR: Winsteps.com.* Retrieved from http://www.winsteps.com/

Messick, S. J. (Ed.). (2013). *Assessment in higher education: Issues of access, quality, student development, and public policy.* Routledge, Taylor and Francis Group.

Mirzaei, A., Hashemian, M., & Tanbakooei, N. (2012). Do Different Stakeholders' Actions Transform or Perpetuate Deleterious High-Stakes Testing Impacts in Iran?. . *The 1st Conference on Language Learning & Teaching: An Interdisciplinary Approach (LLT –IA)*. https://www.sid.ir/en/Seminar/ViewPaper.aspx?ID=24946

Mohammad, S., Amirian, R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting Gender DIF with an English Proficiency Test in EFL Context. *Iranian Journal of Language Testing*, *4*(1), 187–203.

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32(2),* 265–281.

Pae, H. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. *Pearson Education Ltd.*

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2012). Non-Graphical Solutions for Cattell's Scree Test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 23–29. https://doi.org/10.1027/1614-2241/a000051

Rasch Measurement Forum. (2017). Retrieved from http://raschforum.boards.net/.

Rezaee, A. A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji.* 56. 89-108.

Rezai-Rashti, G., & Moghadam, V. (2011). Women and higher education in Iran: What are the implications for employment and the ''marriage market''? *International Review of Education, 57,* 419–441.

Roever, C., & McNamara, T. (2006). Language Testing: The Social Dimension. *International Journal of Applied Linguistics*, *16*(2). https://doi.org/10.1111/j.1473-4192.2006.00117.x

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language testing, 9*(1), 12-29.

Sadeghi, S. (2014). High-stake Test Preparation Courses: Washback in Accountability Contexts. *Journal of Education & Human Developmrnt*, *3*(1), 17–26.

Salehi, M. & Tayebi, A. (2012). Differential item functioning in terms of gender in reading comprehension subtest of a high-stakes test. *Iranian Journal of Applied Language Studies. 4*(1). 135-168.

Salehi, H. & Yunus, M.M., (2012a). The Washback Effect of the Iranian Universities Entrance Exam: Teachers' Insights. *GEMA Online™ Journal of Language Studies. 12*(2),. 609- 628.

Salehi, H. & Yunus, M.M., (2012b). University Entrance Exam in Iran: A bridge or a dam. *Journal of Applied Sciences Research,* 8(2): 1005-1008, 2012. ISSN 1819-544X

Scheuneman, J. D., & Bleistein, C. A. (1989). A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, *2*(3), 255–275. https://doi.org/10.1207/s15324818ame0203_6

Song, X., & He, L. (2015). The Effect of a National Education Policy on Language Test Performance: A Fairness Perspective. *Language Testing in Asia*, *5*(1), 1–14. https://doi.org/10.1186/s40468-014-0011-z

Spolsky, B., & Bachman, L. F. (1991). Fundamental Considerations in Language Testing. *The Modern Language Journal*, *75*(4). https://doi.org/10.2307/329499

Tae, P. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, *32,* 265-281.

Tahmasbi, S., & Yamini, M.. (2012). Teachers' Interpretations and Power in a High-Stakes Test: A CLA Perspective. *English Linguistics Research, 1*(2), 53. https://doi.org/10.5430/elr.v1n2p53.

Terry, R. M., Genesee, F., & Upshur, J. A. (1998). Class-Room-Based Evaluation in Second Language Education. *The Modern Language Journal*, *82*(1). https://doi.org/10.2307/328719

The Glossary of Education Reform. (2014). *11 Ways to Improve School Communications and Community Engagement*. https://www.edglossary.org/school-communications/

Wiberg, M. (2007). Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test : A Theoretic

Comparison of Methods. In *Educational Measurement, technical report N. 2.*

Xi, X. (2010) How do we go about investigating test fairness? *Language Testing,* 27(2), 147-170.

Appendix

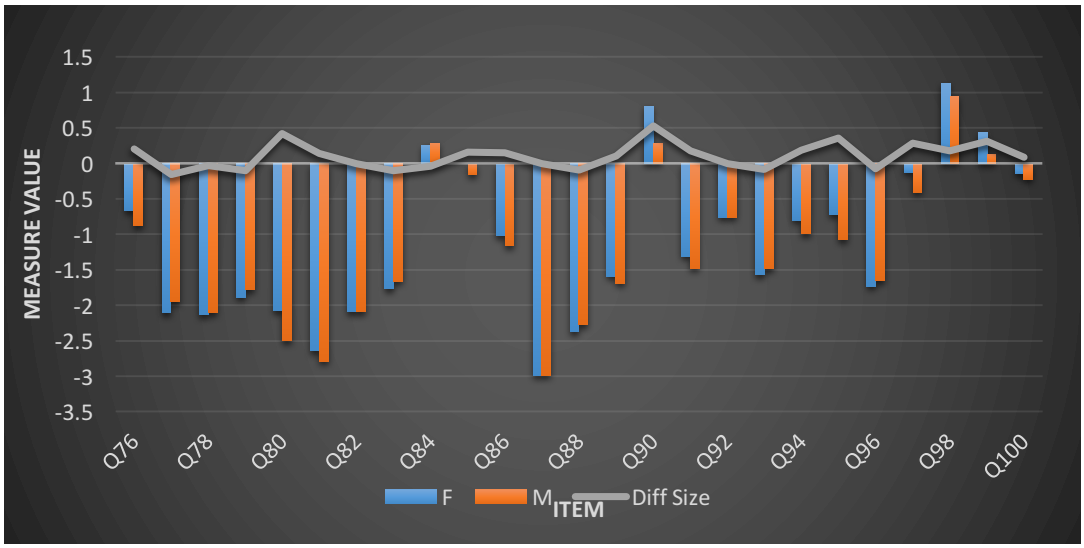

Figure a. DIF size for general questions across gender groups

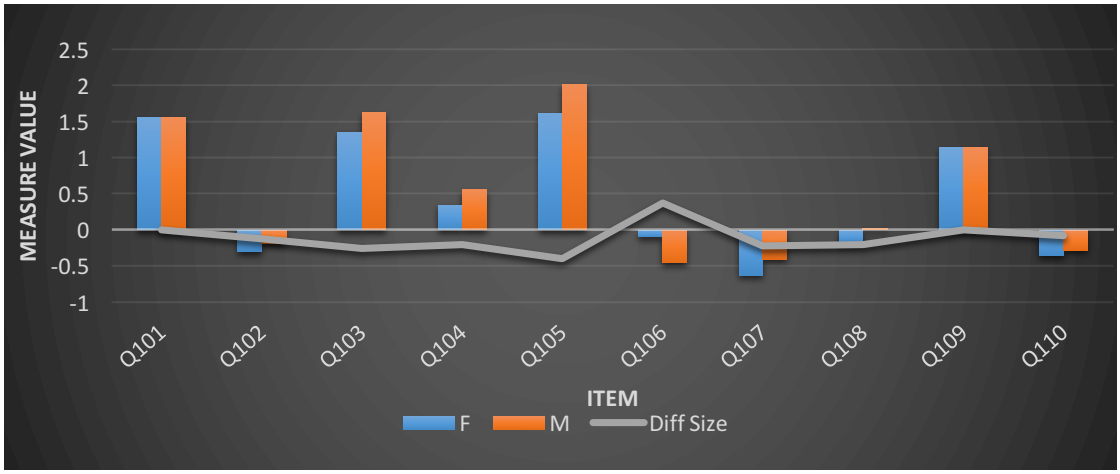

Figure b. DIF size for grammar items across gender groups

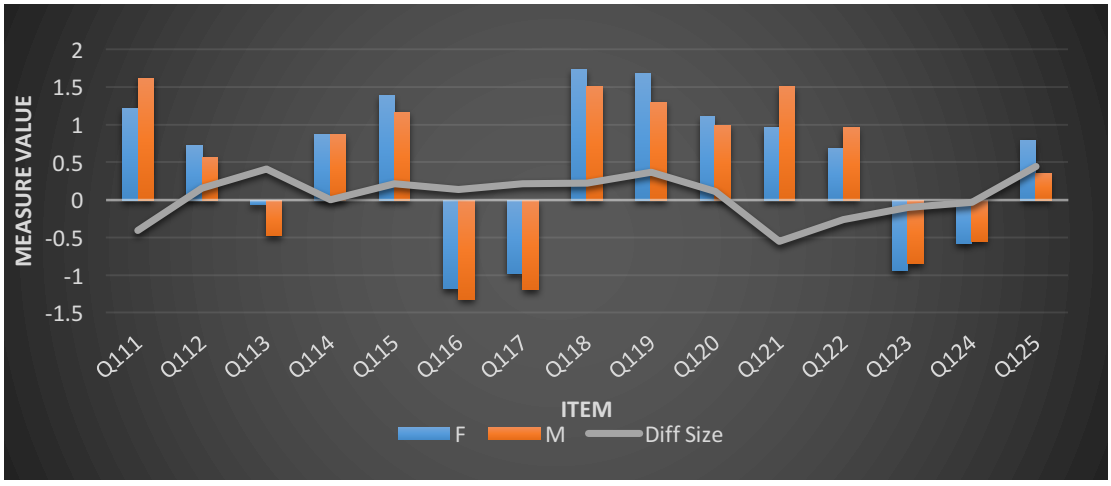

Figure c. DIF size for vocabulary items across gender groups

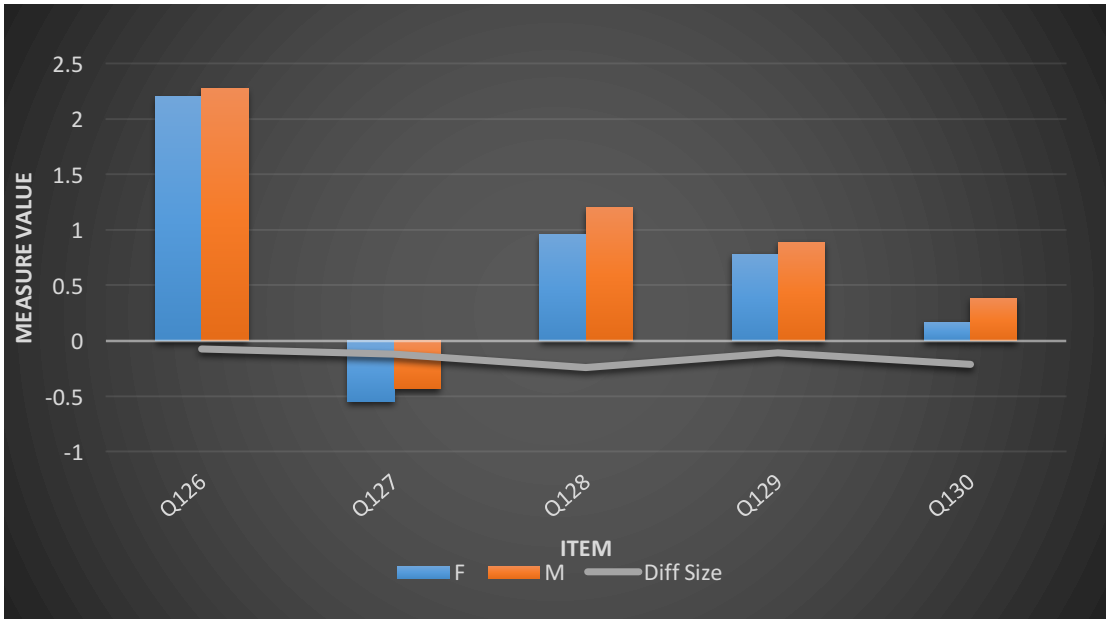

Figure d. DIF size for sentence structure items across gender groups

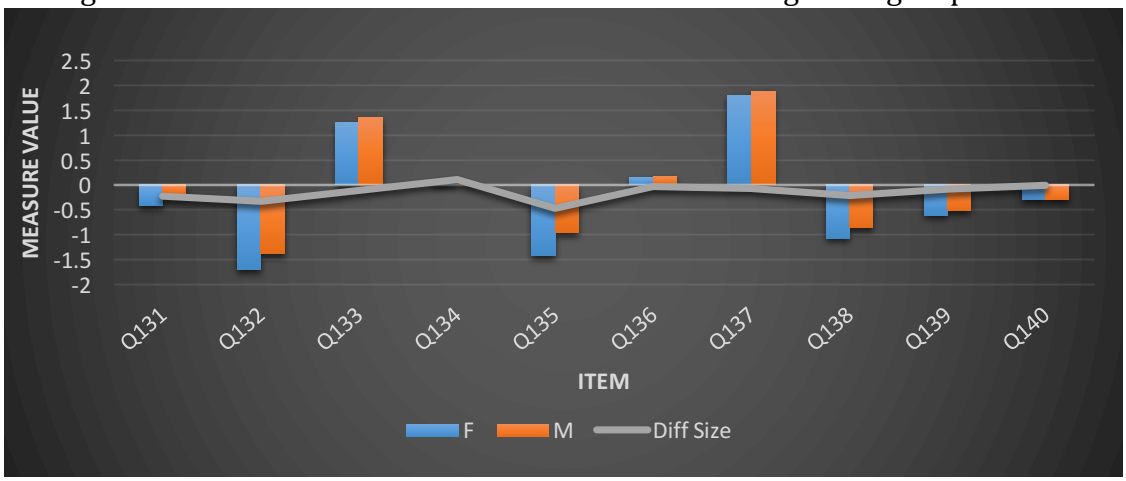

Figure e. DIF size for language function items across gender groups

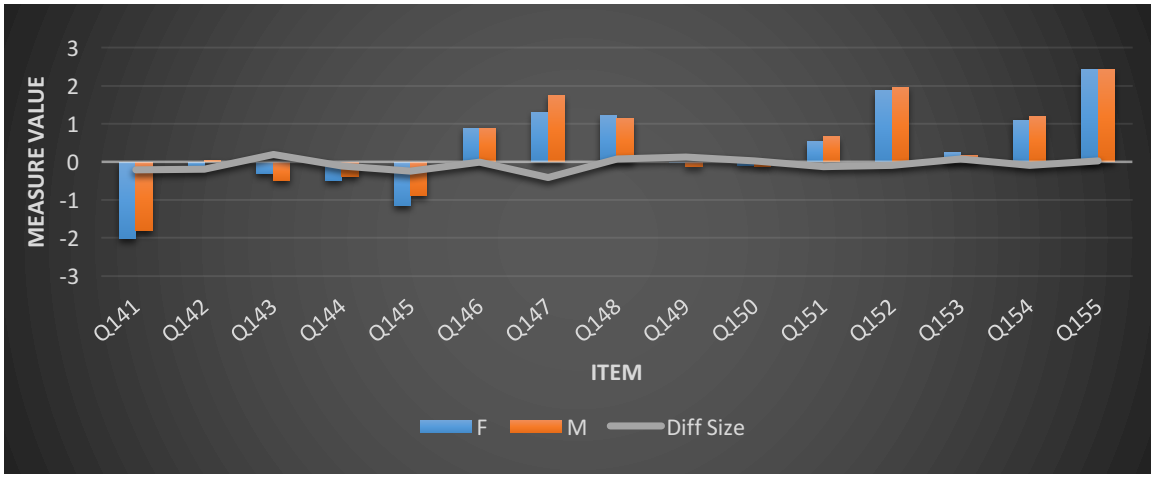

Figure f. DIF size for cloze test items across gender groups