Language Testing in Asia

RESEARCH Open Access

Reading comprehension test fairness across gender and mode of learning: insights from IRT-based differential item functioning analysis

Elahe Moradi, Zargham Ghabanchi and Reza Pishghadam*

*Correspondence: pishghadam@um.ac.ir

Department of English, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

Given the significance of the test fairness, this study aimed to investigate a reading comprehension test for evidence of differential item functioning (DIF) based on English as a Foreign Language (EFL) learners' gender and their mode of learning (conventional vs. distance learning). To this end, 514 EFL learners were asked to take a 30-item multiple-choice reading comprehension test. After establishing the unidimensionality and local independence of the data as prerequisites to DIF analyses, Rasch model-based DIF analysis was conducted across learners' gender and their mode of learning. The results showed that there were two gender-DIF items which functioned differentially in favor of female respondents. Also, DIF analysis in terms of EFL learners' mode of learning revealed that there were no DIF items across the two target groups indicating that the reading comprehension test functioned the same way for learners who enjoyed conventional learning compared to those who experienced distance and self-directed learning. In the end, the findings were discussed and implications were provided.

Keywords: Differential item functioning (DIF), Gender, Mode of learning, Rasch model, Reading comprehension test

Introduction

Literature on test fairness indicated that fairness has had different conceptualizations in its relatively short history (Moghadam & Nasrizadeh, 2020). Test fairness is broadly defined as the equitable treatment of test takers, the lack of bias in measurement, and justifiable uses and interpretations of test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Although different conceptual frameworks of test fairness have incorporated broader concepts such as the socioeconomic aspects of tests or the validity in test fairness (McNamara et al., 2019; Xi, 2010), in a narrow sense, fairness is concerned with the consistency of test functions, most specifically whether the background characteristics of test takers influence their performance on language tests (McNamara & Ryan, 2011). The researchers of this study adopted McNamara and Ryan's (2011) definition of

test fairness in the current study, which refers to "the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test-takers and the adequacy of the representation of the construct in test materials and procedures" (p. 163). To investigate test fairness from this point of view, differential item functioning (DIF) investigation across sub-groups at item level has been adopted (Ferne & Rupp, 2007; Zumbo et al., 2015; Zhu and Aryadoust, 2019). DIF refers to the extent to which test items function differently for different subgroups of test takers with the same ability level because the test is measuring off-trait characteristics such as gender, mother tongue, or academic background (Banerjee & Papageorgiou, 2016). DIF analysis has been used to provide evidence as to whether the test actually measures the ability intended to be measured and does not put any groups in a favorable or unfavorable position, thus shedding light on the sources of construct-irrelevant variance in test scores (Messick, 1989).

The Standards for Educational and Psychological Testing (AERA) state that all assessments should be screened for items that may exhibit DIF in order to ensure fairness in testing. So researchers and test users should be cautious with test bias analysis and how it may affect test performance. Conducting a fair test is an important issue in social research as well. Many important decisions are made based on the results of tests administered under different conditions in the fields of educational and psychological testing. Inaccurate inferences are often made if the property of measurement invariance is not assessed across these conditions (Makransky & Glas, 2013).

Test bias can cause serious effects, both for the individuals and for the society in general. Hence, it is crucial to assure that educational tests are unbiased and the scores obtained from them are not unduly affected by construct-irrelevant factors, so that the test does not unfairly advantage or disadvantage one group of the examinees over the others. What this postulates is that sources of bias need to be identified and dealt with in test design and score interpretation to minimize bias and its negative consequences.

A probe into the related literature depicted that numerous DIF studies have been conducted to identify group differences in testees' performance and to examine whether test items are invariant across members of different subgroups in the context of DIF. The grouping has been made in terms of gender (e.g., Lawrence & Curley, 1989; Maller, 2001; MacIntyre et al., 2002; Breland et al., 2004, Lumley and O'Sullivan, 2005; Aryadoust et al., 2011; Chubbuck et al., 2016), ethnicity (e.g., Sehmitt & Dorans, 1990; Oliveri et al., 2018), academic backgrounds (e.g., Pae, 2004), linguistic backgrounds (e.g., Chen & Henning, 1985; Ryan & Bachman, 1992), topic knowledge (e.g., Khabbazbashi, 2017), native language (Allalouf & Abramzon, 2008), and disability status (e.g., Maller, 1997).

Despite the fact that several research studies have investigated various sources of test bias such as gender, ethnic, language background, academic background, and disability status, no study, thus far, has investigated DIF across learners' mode of learning to uncover whether conventional learning at regular universities or self-directed learning at distance universities can distort the measurement invariance in EFL learners' reading comprehension test performance. Hence, the significance of the current investigation mainly lies in its attempt to deal with the role of EFL learners' mode of learning (conventional vs. distance) along with their gender in bringing about differential item functioning in reading comprehension test performance. To the best of our knowledge, this is the

first study scrutinizing the role of students' "learning mode" in provoking bias in reading comprehension test performance. Employing DIF analysis, this study investigates whether the reading comprehension test functions fairly and in a similar way among EFL conventional and distance learners.

As the related literature noted, there exist several DIF analysis techniques such as the Rasch models, the Mantel-Haenzel method, the multi-dimensional item response theory models, the logistic regression method, and the simultaneous item bias test (SIBTEST). The present study employed Rasch model-based DIF analysis method to examine differential item functioning across respondents' gender and their mode of learning. The advantage of the Rasch model over other DIF detection methods is that parameters in Rasch models are sample-independent. The Rasch model computes item difficulties and person abilities on a strictly linear scale based on item and person raw scores without requirements such as complete data and normal distributions. The effect is to limit the impact of guessing and item discrimination. This makes the Rasch model more convenient in terms of utility, inference, and explanation compared to other IRT models.

Furthermore, to conduct the current study, the authors adopted the framework of the "second DIF generation" proposed by Zumbo (2007). In other words, the authors of this study focused on the investigation of differential item functioning and unidimensionality analysis. Although DIF analysis has developed noticeably over the past decade, the related literature still suffers from lacking a firm theory on the means of DIF detection (Zumbo, 2007). There exist two approaches on DIF investigation including exploratory approach, in which the researchers conduct a post hoc content analysis of the items indicating DIF or confirmatory approach, in which the researchers analyze test items to generate hypothesis using DIF detection methods. According to Ferne and Rupp (2007), most researchers preferred to use exploratory DIF analysis in the field of language assessment, because they believed that exploratory analysis can lead to new theories of DIF. For this reason, the researchers of this study took an exploratory approach to investigate DIF across EFL learners' gender and their mode of learning in an effort to close the existing gap in the related literature.

The objectives of this research were to investigate:

1. Whether reading comprehension test meets the requirements of local independence and unidimensionality as pre-conditions for Rasch-based DIF analysis
2. Whether there exists evidence of gender-based DIF in the reading comprehension test
3. Whether there exists evidence of DIF in the reading comprehension test in terms of EFL learners' mode of learning (conventional vs. distance learning)

Literature review

The Rasch-based DIF analysis

The Rasch-based DIF analysis exists when different groups of test takers of the same ability level have significantly different chances of answering a test item because the test interacts with off-trait characteristics such as test takers' gender or mother tongue (Engelhard, 2012). DIF analysis has been applied to investigate the validity of the uses

and interpretations of test scores, as the presence of DIF might indicate the undesirable effect of some factors that are not relevant to the construct under assessment. The presence of DIF has also been regarded as a violation or attenuation of test fairness by McNamara and Ryan (2011) who alluded to Messick's (1989) validity framework to argue that investigating fairness would provide the evidential basis for the better utilization and interpretation of test scores.

From a psychometric perspective, there are a number of factors which can adversely affect test fairness and which the designers of high-stakes tests should monitor closely such as gender, age, language background, ethnicity, and academic experience. Research has shown that these factors can yield significant DIF and put certain groups to an advantage or disadvantage. To exert sizeable impacts on test takers' measured performances, the detected DIF must be statistically significant ($p < .05$) and substantive (Bond & Fox, 2015; Linacre, 2018); in this case, DIF would indicate that test scores are confounded by unintended factors beyond the construct which jeopardize test fairness (Messick, 1996; Wright & Stone, 1999).

In language assessment, even though gender is regarded as one of the "more stable" test taker characteristics, which means that it is not likely to elicit differential treatment on test takers compared to the "less stable" characteristics such as strategies, skills, and motivations (Alderson, 2000, p. 56), a number of studies have reported DIF across female and male test takers (Carlton & Harris, 1992; Lawrence and Curley, 1989; Ryan & Bachman, 1992; Pae, 2004; Zhu and Aryadoust, 2019). This suggests that the constructs under investigation were measured in different ways for females and males even if they match on the overall ability level (Camilli & Shepard, 1994).

Recently, several studies have been conducted in the domain of test fairness. Karami (2013) examined the presence of gender differential performance on the University of Tehran English Proficiency Test. To this end, the performance of 2343 test takers was analyzed. The results of data analysis revealed that maleness or femaleness did not make any contributions to score variance. In other words, it was shown that the test is free of gender bias and is highly reliable.

Khabbazbashi (2017) conducted a study to explore the effect of topic and background knowledge of topic on a speaking test performance. The data sources in the study were eighty one non-native speakers of English who were assigned 10 topics across three task types. The results of the study showed that, although with little practical significance, different levels of background knowledge of the topic systematically influenced the participants' speaking performance.

Zhu and Aryadoust (2019) examined test fairness across gender in a computerized reading test. Two quantitative approaches to investigate test fairness, the Rasch-based differential item functioning (DIF) detection method and a measurement invariance technique called multiple indicators, multiple causes (MIMIC), were used to investigate the Pearson Test of English (PTE) Academic Reading Test for the evidence of DIF in terms of gender. To this end, the researchers selected 783 participants and conducted the Rasch partial credit model to flag DIF items. The results showed no statistically significant DIF in terms of gender and, in the same way, the multiple indicators multiple analyses proved measurement invariance of the test.

In another study, Zhu and Aryadoust (2020) conducted research to evaluate the fairness of the Pearson Test of English (PTE) Academic Reading Test to examine whether language background can bring about differential item functioning across Indo-European and non-Indo European language families. Using partial credit Rasch model, they carried out uniform (UDIF) and non-uniform (NUDIF) analyses on 783 international test-takers. The results indicated no statistically significant uniform DIF, but three NUDIF item out of 10 items across the language families. According to the findings, a mother tongue advantage was not observed. The content analysis of non-uniform DIF items indicated that the decrease of mother tongue advantage for Indo European high ability groups and lucky guesses for low ability group contributed to differential item functioning.

Similarly, Rashvand and Ahangari (2022) investigated DIF in the MSRT (Ministry of Science, Research and Technology) test items. To this end, 200 pre-intermediate and intermediate EFL Iranian learners with the age range of 25 to 32 in two different fields of study (100 humanities and 100 sciences) were randomly selected for the analysis. The researchers examined DIF across students' majors using item response theory likelihood ratio approach to identify items displaying DIF. Their findings identified 15 DIF items favoring science test-takers. The results demonstrated that the test was statistically easier for the science test-takers at 0.05 level.

Bordbar (2021) used Rasch-based DIF analysis as a technique to investigate test fairness and detected gender-biased items in the National University Entrance Exam for Foreign Languages (NUEEFL). The researcher concluded that the test scores were not free of construct-irrelevant variance, and certain inaccurate items were modified following the fit statistics guidelines. In the same line, Butler and Iino (2021) conducted a study in Japan to examine the notion of test fairness in College Entrance Exam (CEE) in terms of students' socio-economic status. Their findings corroborated that the exam structurally worked against students with lower socio-economic status and introduced this variable as a source of test unfairness.

Among different factors which can lead to forming biases in individuals' test performances, no study, thus far, has investigated the role of "learning mode" as a source of test bias which in turn can differentially impact test taking performances. For this reason, the researchers of this study aspired to investigate the role of EFL learners' mode of learning along with their gender as a source of reading comprehension test unfairness.

Methodology

Participants

The participant of this study included 100 male and 414 female intermediate EFL learners from three universities in Mashhad, Iran. The participants were selected based on convenience or opportunity sampling and their ages ranged between 18 and 45 ($mean_{age}$= 25.63, SD= 6.43). Also, 287 individuals out of these 514 participants experienced distance learning (distance university students), and the other ones experienced conventional learning (conventional university students) as their mode of learning. The participants were ensured about the confidential nature of this study and they were informed that participation was completely voluntary.

It should be noted that distance education is flexible education where learners are free from the constraints of time, pace, and place of the study. In distance universities, the number of classes is approximately one third or even one fourth the number of classes in conventional universities and the classes are not obligatory to attend. Also, students mostly experience self-study and self-directed learning in such universities. Thus, students who are working or those who are from more remote regions and prefer to remain at their own city for their university studies select this mode of learning to pursue their higher education.

Instruments

First, to ensure the homogeneity of the participants in terms of their English language proficiency level, the Oxford Quick Placement Test was administered among EFL learners. Then, in order to collect the required data, a 30-item multiple-choice reading comprehension test was employed which comprised three passages. The reading comprehension test was chosen from the reading component of Longman Complete Course for the TOEFL Test (Philips, 2005).

Results

To conduct DIF analysis, two major analyses should be initially done on the test data. First, the exploration of descriptive statistics, item difficulty measures, fit of the data into the Rasch model, and estimating reliability. Second, the investigation of uni-dimensionality and the degree of local independence of the items.

Descriptive statistics for the data set were calculated including mean, standard deviation, and skewness and kurtosis coefficients, using SPSS for Windows, Version 24. Furthermore, to analyze the collected data for detecting differential item functioning (DIF), Winsteps 3.74.0 Software was employed.

As can be seen in Table 1, skewness and kurtosis coefficients fall between -2 and $+2$ in all items, indicating univariate normality. Item 24 with the mean of .19 and the standard deviation of 0.39 was the most difficult item in which the participants were asked to make an inference based on the information provided by the passage. Item 13 with the mean of .71 and the standard deviation of .45 was the easiest one in which students were asked to find the reference for a pronoun.

Rasch reliability analysis

The item reliability coefficient is .97> (.9), meaning that only 3% of the variability in item is attributable to error. The difficulty strata or the number of different abilities a test can determine is shown by the item separation index. Item separation value for the reading comprehension test is 8.25 (>3) which is a satisfactory index. Thus, the representativeness of the items of the test is verified.

Unidimensionality and local independence

Rasch measurement was used to test both the unidimensionality and the local independence, as prerequisites to DIF analysis. First, to detect the unidimensionality of the test, principal component analysis (PCA) of standardized residuals was investigated.

Table 1 Descriptive statistics

Item	Mean	SD	Skewness	Kurtosis
1	.66	.47	−.69	−1.53
2	.52	.50	−.09	−1.99
3	.61	.48	−.45	−1.79
4	.52	.50	−.09	−1.99
5	.33	.46	.74	−1.45
6	.34	.47	.69	−1.52
7	.33	.46	.74	−1.45
8	.43	.49	.28	−1.92
9	.50	.50	.00	−2.00
10	.37	.48	.54	−1.70
11	.38	.48	.48	−1.77
12	.37	.48	.52	−1.73
13	.71	.45	−.93	−1.14
14	.43	.49	.30	−1.91
15	.39	.48	.43	−1.81
16	.35	.47	.65	−1.58
17	.29	.45	.91	−1.17
18	.63	.48	−.53	−1.71
19	.49	.50	.05	−2.0
20	.23	.41	1.31	.26
21	.27	.44	1.04	.91
22	.36	.48	.56	−1.68
23	..37	.48	.56	−1.69
24	.19	.39	1.58	.52
25	.51	.50	−.04	−2.00
26	.48	.50	.06	−2.00
27	.65	.47	−.61	−1.62
28	.52	.50	−.06	−2.00
29	.30	.45	.89	−1.19
30	.35	.47	.64	−1.59

Table 2 Dimensionality output

	Empirical			Modeled
Total raw variance in observations	36.6	100.0%		100.0%
Raw variance explained by measures	6.6	17.9%		17.7%
Raw variance explained by persons	2	5.5%		5.5%
Raw Variance explained by items	4.5	12.4%		12.2%
Raw unexplained variance (total)	30.0	82.1%	100.0%	82.3%
Unexplained variance in 1st contrast	2.1	5.6%	6.9%	

PCA of the standardized residuals (Table 2) showed that the Rasch dimension is as big as 6.6 items which explains 17.9% of the variance; 12.4% are explained by item measures and 5.5% are explained by person measures.

The eigenvalue of the first observed contrast was 2.1 (rounded to 2), which supports the presumption that the data set is unidimensional meaning that, targeted on the same underlying construct.

Next, to examine the local independence, the residuals correlations were checked. According to (Linacre, 2018), correlations above .70 are evidence of local dependence. The findings showed that all the observed correlations were less than .70.

If some item residuals cluster together and construct a component, it can be considered as an indication of high common variance among them. As the principal component analysis (PCA) of standardized residuals in Fig. 1 revealed, the residuals do not form distinguishable patterns or clusters.

Fit of the data to the latent trait model

Table 3 presents Rasch measurement results, including difficulty measures in logits and fit indices.

Infit and Outfit MNSQ (mean-square) statistics were presented in the table. The expected value for MNSQ is 1 (Linacre, 2012); however, the range of .7 to 1.3 or 1.4 is recommended (Bond & Fox, 2015; Linacre, 1999).

Examining the infit and outfit mean-squares associated with the responses to each item of the test indicated that there were no misfitting items and all the items are within the acceptable limit; thus, all the items fit the Rasch model. Furthermore many items show MNSQ fit indices near 1, showing a lack of erratic response patterns in the data.

Item-person map

The item-person map in Fig. 2 indicates the relationship between person performance estimates and item difficulty measures along a single line in logits. The conformity of the bulk of items on the right to the bulk of persons on the left is an indication that the

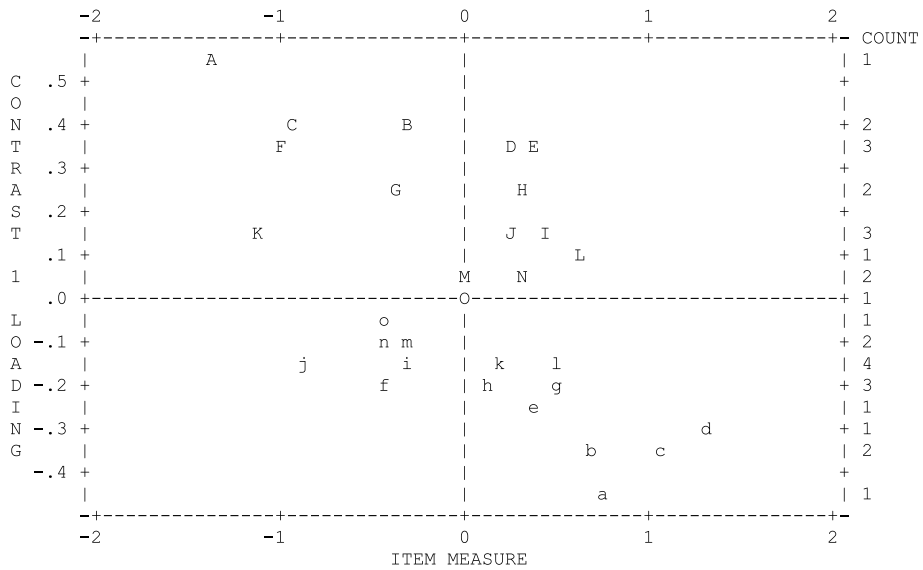

Fig. 1 Principal component analysis of linearized Rasch residuals of the data

Table 3 Item measures and fit statistics for the items of reading comprehension test

Item	Measure	Error	Infit MNSQ	Outfit MNSQ
24	1.29	.12	1.02	1.16
20	1.06	.11	1.22	1.36
21	.79	.11	1.32	1.46
29	.70	.11	.94	0.95
17	.62	.10	1.20	1.23
5	.43	.10	1.08	1.15
7	.41	.12	.98	1.03
30	.41	.10	.97	.97
6	.38	.10	.97	.98
16	.36	.10	.87	.83
23	.31	.10	.87	.84
22	.29	.10	.86	.82
10	.24	.10	.94	.92
12	.21	.10	.95	.92
15	.18	.10	1.15	1.15
11	.15	.10	1.12	1.17
8	−.1	.10	.94	.92
14	−.4	.10	1.06	1.05
26	−.15	.90	.89	.86
19	−.27	.90	1.06	1.08
28	−.31	.90	.96	.94
25	−.35	.90	.88	.86
9	−.39	.90	1.07	1.11
4	−.50	.90	1.02	1.01
2	−.52	.90	1.03	1.03
3	−.91	.10	1.04	1.01
27	−.94	.10	0.89	.86
18	−.96	.10	0.92	.87
1	−1.12	.10	0.96	.90
13	−1.36	.10	0.85	.77

difficulty of the test items aptly addresses the respondents. In other words, items are representative for estimating the participants' ability levels.

The map shows that test items cover a relatively wide range of difficulty, from −1.35 logits (Item 13) to +1.28 logits (Item 24). As the map illustrates, there are some low-ability test takers with no test items in their ability levels.

After establishing the two preconditions for DIF analysis, i.e., unidimensionality, and local item independence, the researchers of the current study conducted DIF analysis across participants' gender and their mode of learning to examine whether these variables may lead to measurement invariance and forming bias in EFL learners' reading comprehension test performance.

Differential item functioning (DIF) analysis across gender

Table 4 presents the item difficulties for each group, followed by their standard errors. The higher the DIF measure, the more difficult the item is. The next column provides the difficulty contrast between the two groups. If the DIF contrast has a positive value,

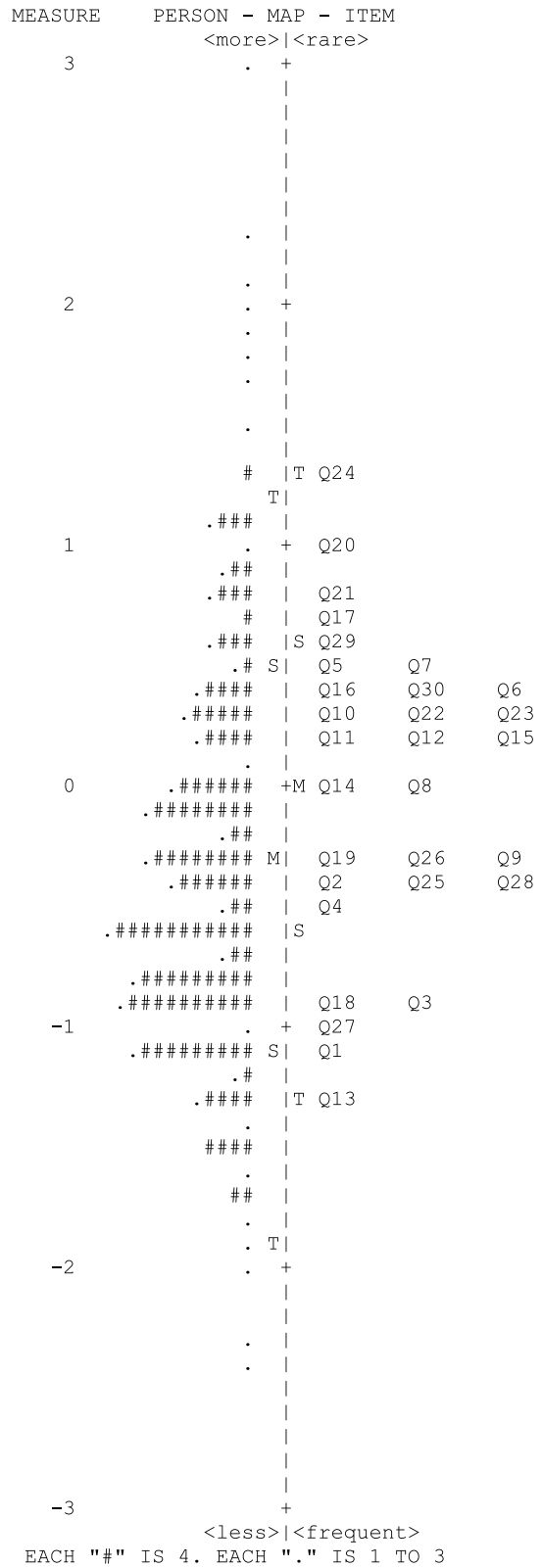

Fig. 2 Item-person map

Table 4 DIF statistics for the reading comprehension test in terms of learners' gender

Item	Female Difficulty (S.E.)	Male Difficulty (S. E.)	Difficulty contrast (S. E.)	Weltch *t*	Weltch Prob
1	−1.06(.11)	−1.27(.24)	.20(.26)	.76	.44
2	−.36(.10)	−.79(.22)	.42(.25)	1.71	.08
3	−.85(.11)	−.85(.23)	.00(.25)	.00	1.0
4	−.52(.11)	−.19(.22)	−.32(.24)	−1.35	.18
5	.51(.11)	.38(.22)	.13(.25)	.50	.61
6	.41(.11)	.37(.22)	.04(.25)	.18	.86
7	.52(.11)	.33(.22)	.19(.25)	.76	.44
8	−.07(.11)	.14(.22)	−.21(.25)	−.86	.39
9	−.39(.11)	−.15(.22)	−.24(.24)	−1.00	.31
10	.26(.11)	.34(.22)	−.08(.25)	−.32	.74
11	.20(.11)	.23(.22)	−.03(.25)	−.11	.90
12	.27(.11)	.11(.22)	.16(.25)	.65	.51
13	−1.33(.11)	−1.46(.25)	.14(.27)	.50	.61
14	−.01(.11)	.02(.22)	−.03(.24)	−.12	.90
15	.04(.11)	.63(.23)	−.60(.26)	−2.33	.02
16	.44(.11)	.14(.22)	.30(.25)	1.22	.22
17	.64(.12)	.74(.24)	−.11(.26)	−.41	.68
18	−.90(.11)	−1.12(.23)	.21(.26)	.83	.40
19	−.26(.11)	−.39(.22)	.14(.25)	.56	.57
20	.98(.13)	1.27(.26)	−.28(.29)	−.97	.33
21	.87(.12)	.46(.23)	.41(.26)	1.58	.11
22	.33(.11)	.10(.22)	.23(.25)	.91	.36
23	.32(.11)	.16(.22)	.15(.25)	.60	54
24	1.35(.14)	1.05(.25)	.30(.29)	1.06	.29
25	−.36(.11)	−.60(.22)	.24(.25)	.98	.32
26	−.33(.11)	−.09 (.23)	−.24(.25)	−.95	.34
27	−1.07(.11)	−.83(.23)	−.24(.25)	−.96	.34
28	−.54(.11)	.09(.22)	−.63(.25)	−2.53	.01
29	.60(.12)	.71(.24)	−.11(.27)	−.42	.67
30	.37(.11)	.37(.23)	.00(.26)	.00	1.0

it is indicated that males are favored and if the value is negative, females are advantaged. However, DIF should be big enough to be noticeable. Usually the difference of .50 logits is recommended (Linacre, 2012). Also, the table presents a Welch *t* and a Welch probability. For statistically significant DIF on an item, usually a probability of less than .05 is needed. Significant DIF indices imply that test scores do not represent only the intended latent variable and there exists an unintended extra dimension (Wright & Stone, 1988).

As the table suggests, items (15 and 28) showed significant gender-based DIF. The difficulty of item 15 is .04 for female and .63 for male respondents. Since .63 is higher than .04, this item is more difficult for male students. This contrast is significant (DIF contrast= −.60 >.50, and *p*-value=.02 <.05).

As for item 28, the difficulty level of this item is −.54 and .09 respectively for female and male students indicating that this item is more difficult for male respondents. Since (the difficulty contrast = −.63 > .50, and *p*-value =.01 <.05); it can be concluded that female students significantly outperformed males on this item.

Analyzing the content of these two gender-based DIF items revealed that items (15 and 28) which were related to "making inferences and drawing conclusions" were proved to have DIF, meaning that in such items there were more significant differences between the performance of male and female examinees, favoring the female ones.

Figure 3 shows the DIF plot which was drawn based on the DIF measures of each individual test item in terms of EFL learners' gender.

Differential item functioning (DIF) analysis across mode of learning

To examine differential item functioning (DIF) across EFL learners' mode of learning, the participants were assigned code of 1 (Distance Learners) and code of 2 (Conventional Learners). Then, DIF analysis was conducted across these two target groups. As Table 5 indicated, no items flagged DIF concerning EFL learner's mode of learning (no DIF contrast is above .50 logits). This result suggested that EFL learners' mode of learning had no significant impact on their reading comprehension test performance and all the test items functioned the same way for the two target groups, i.e., "Distance learners" who have learned mostly via self-study and self-directed learning and "Conventional learners" who enjoyed attending more classes and had more face to face learning with their teachers.

Figure 4 shows the DIF plot which was drawn based on the DIF measures of each individual item in terms of EFL learners' mode of learning.

Discussion

According to the joint Standards, American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME), no test should have bias; it should be fair enough so as to provide a standardized testing procedure to guarantee reasonable treatment of examinees, and have equality of testing outcomes for subgroups of examinees by race, gender, background knowledge, etc. (Jiao & Chen, 2013). Research has shown that these factors can yield significant DIF and put certain groups to an advantage or disadvantage status. As a

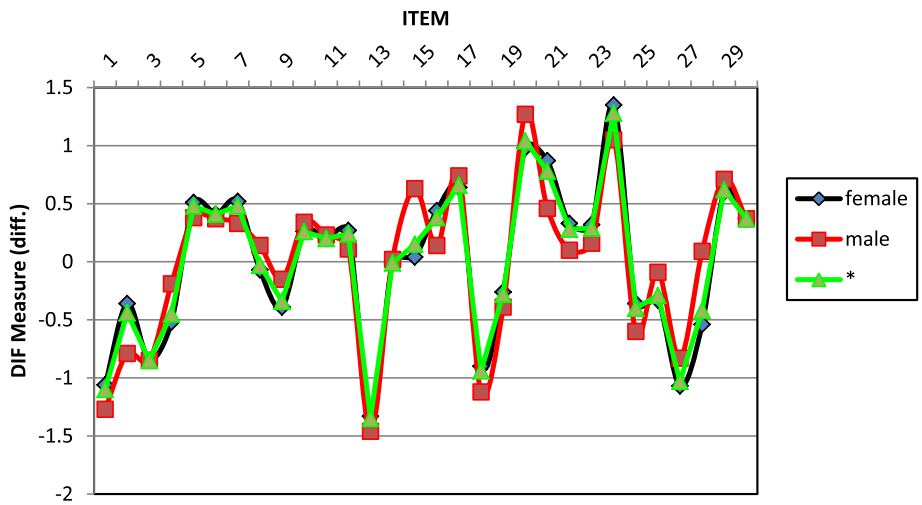

Fig. 3 Differential item functioning (DIF) for EFL learners' gender

Table 5 DIF statistics for the reading comprehension test in terms of learners' mode of learning

Item	Code1 Difficulty (S.E.)	Code2 Difficulty (S. E.)	Difficulty contrast (S. E.)	Weltch t	Weltch Prob
1	.47(.06)	.47(.06)	.00(.08)	.00	1.00
2	.56(.06)	.65(.06)	−.09(.09)	−1.07	.28
3	.25(.06)	.22(.06)	.02(.08)	.25	.80
4	−.26(.06)	−.31(.07)	.05(.09)	.58	.56
5	−.05(.06)	−.11(.06)	.06(.09)	.71	.47
6	.039(.06)	.23(.06)	−.20(.08)	−2.32	.02
7	−.01(.06)	−.06(.06)	.05(.09)	.55	.58
8	.08(.06)	.08(.06)	.00(.08)	.00	1.00
9	−.58(.06)	−.51(.07)	−.07(.09)	−.77	.44
10	.56(.06)	.68(.06)	−.12(.09)	−1.41	.16
11	.88(.06)	.88(.07)	.00(.09)	.00	1.00
12	.08(.06)	.08(.06)	.00(.08)	.00	1.00
13	.23(.06)	.23(.06)	.00(.08)	.00	1.00
14	.01(.06)	−.11(.06)	.12(.09)	1.39	.16
15	−.40(.06)	−.53(.07)	.13(.09)	1.42	.15
16	.44(.06)	.44(.06)	.00(.08)	.00	1.00
17	−.41(.06)	−.50(.07)	.09(.09)	.94	.34
18	−.20(.06)	−.20(.07)	.00(.09)	.00	1.00
19	−.27(.06)	−.27(.07)	.00(.09)	.00	1.00
20	.44(.06)	.44(.06)	.00(.08)	.00	1.00
21	−.13(.06)	−.22(.07)	.09(.09)	1.01	.31
22	−.60(.06)	−.60(.07)	.00(.10)	.00	1.00
23	−.17(.06)	−.17(.07)	.00(.09)	.00	1.00
24	−.69(.06)	−.69(.07)	.00(.10)	.00	1.00
25	−.35(.06)	−.35(.07)	.00(.09)	.00	1.00
26	.18(.06)	.18(.06)	.00(.08)	.00	1.00
27	.23(.06)	.19(.06)	.05(.08)	.55	.58
28	−.22(.06)	−.08(.06)	.14(.09)	−1.60	.11
29	−.06(.06)	−.06(.06)	.00(.09)	.00	1.00
30	−.02(.06)	−.02(.06)	.00(.09)	.00	1.00

consequence, a scrutinized attention must be paid in analyzing test results so that what are being measured will be test takers' true abilities. This can be achieved by diminishing irrelevant factors and increasing the effect of the ability factor of individuals taking the test. Accordingly, this study set out to examine DIF caused by EFL learners' gender and their mode of learning in a reading comprehension test driven from the reading component of Longman Complete Course for the TOEFL Test, using Rasch model.

As mentioned before, there are two prerequisites for Rasch-based differential item functioning (DIF) analysis, namely unidimensionality and local independence. Unidimensionality refers to a situation in which test scores are not infected by any irrelevant element, and local independence as one of the underlying assumptions of latent variable models assumes that test takers' response to a given test item does not influence their response to another item (Ferne & Rupp, 2007). To explore the unidimensionality and local independence of the test, the results of principal component analysis (PCA) of standardized residuals were investigated and it was revealed that

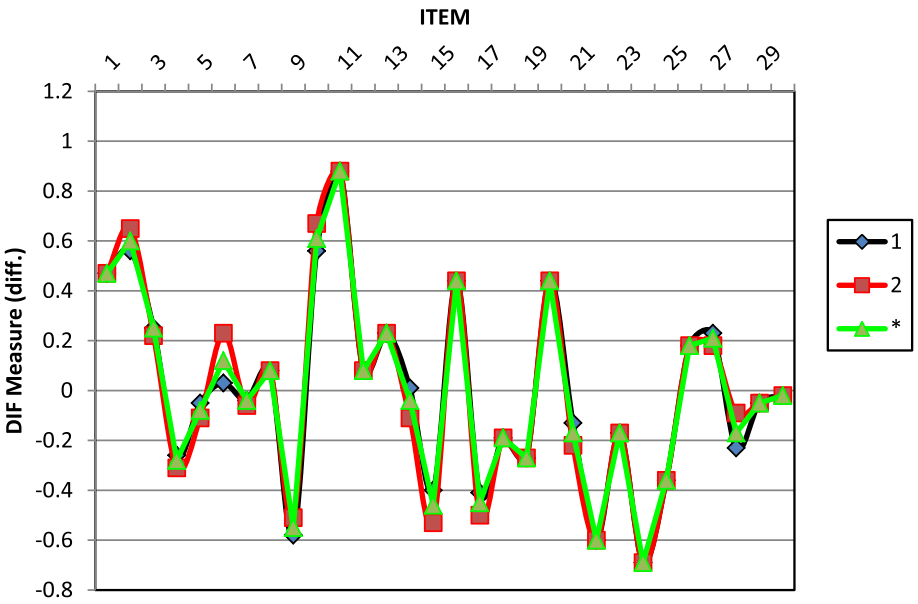

Fig. 4 Differential item functioning (DIF) for EFL learners' mode of learning

the reading comprehension test meets the requirements of unidimensionality and local independence as pre-conditions for Rasch-based DIF analysis. As the data analysis indicated, there were no misfitting items and all the items showed good fit to the Rasch model. Moreover, examining fit statistics demonstrated no erratic patterns in the data, and this diminishes the effect of guessing or carelessness in item responses. The item-person map also demonstrated that the reading comprehension test items were well targeted for measuring the participant' reading test performance.

After establishing unidimensionality and local independence of the reading comprehension test as the perquisites to DIF analysis, Rasch model-based DIF analysis was conducted for identifying gender-based DIF items. As the findings indicated, item 15 (What did the willow bark look like after Stone prepared it for his experiment?) and item 28 (Why are scientists sure that the prehistoric coelacanth was a flesh-eater?) flagged DIF, meaning that these items were easier for female respondents than the male ones (both significant at $p<.05$). Since test item content is a likely cause of observed DIF, content analysis of these items was done by three English language experts and they suggested that females outperformed males on these items which were more related to making inferences and drawing conclusions based on the provided context in the reading comprehension test. Making an inference involves using what an individual knows to make a guess about what s/he does not know. Readers, who make inference more effectively, use the clues in the text along with their own experiences more adequately to guess what is not directly expressed in the text. Here as the findings showed, male learners found it harder to make inference compared to the female ones. To shed more light and delve more deeply into the probable sources of DIF on such items, more pieces of evidence are required from more empirical studies to confirm. Overall, as the gender-DIF results in this study indicated, the findings were in line with some previous studies (e.g., Lawrence and Curley, 1989; van Langen

et al., 2006; Pae, 2012) where gender was introduced as a potential source of bias in reading comprehension tests.

With respect to the third aim of the current study, the result of DIF detection based on EFL learners' mode of learning indicated no significant DIF indices, suggesting that the reading comprehension test items did not function differentially across the two different modes of learning (conventional versus distance learning). Hence, it can be concluded that there was no significant difference between reading comprehension performance of those students who attended distance university and experienced self-directed learning and those who enjoyed conventional learning at regular universities. As was evident from the value of DIF contrasts, all the items of the reading comprehension test functioned in the same way for both conventional and distance learners. The results showed that mode of learning did not bring about any bias in participants' responses to the items of the reading comprehension test. Thus, it can be inferred that distance learning can be as effective as conventional learning and those who prefer a flexible and self-paced learning and tend to be free from the constraints of time and place of study can enjoy the same performance as those who attend more classes and prefer to have more face to face learning.

On the whole, the aim of this study on DIF was to compare the performance of male/female students and conventional/distance learners on reading comprehension test items through an exploratory approach as to understand the weaknesses of the test to have a fair test. The results of this study identify the need for regular evaluation and revision of tests to make the measurement process as precise as possible. Test developers should play critical roles in the assessment process. It is suggested that test designers construct tests free of bias. Although the reading comprehension test utilized in this study was a test driven from TOEFL test as one of the most reliable and valid tests, DIF was still present in the reading questions. This finding reinforces the point that when reading comprehension is involved, the issue of measurement may be confounded by the gender of the participants. Interestingly, the findings of the study confirmed the fairness of the reading comprehension test across learners' mode of learning, meaning that the students' reading performance has not been affected by their learning mode and the way they received input. The results of this study corroborated that reading comprehension test was not biased against those students who were unable to attend campus-based or full time courses. It was shown that the learners who had more personalized learning and study at their own pace could achieve similar reading comprehension test performance compared to the conventional learners at regular universities who enjoyed more interpersonal skills in their classes.

Conclusions

In the present study, an attempt was made to scrutinize a reading comprehension test for the evidence of test unfairness in terms of EFL learners' gender and their mode of learning.

To this end, the role of "learning mode" in flagging DIF was investigated for the first time and the study's findings documented that EFL learners' reading performance was not affected by their mode of learning and this variable did not perform as a probable source of reading comprehension test bias. Thus, test fairness evidence was found for

reading comprehension test among distance and conventional learners. Furthermore, the results indicated that there were two items which showed different behavior patterns in terms of EFL learners' gender; however, further studies are needed to ascertain the underlying causes.

This study as a pioneer in the related area can open up new perspectives for the future studies. In light of the findings of the current study, some advice is presented for test developers. The researchers' main recommendation is for test constructors. They should be aware of the negative effect of DIF on test validity and test results so as to design and construct tests which do not advantage any group by considering as various factors as possible. It is also recommended that the items displaying significant DIF be analyzed and revised. It is necessary to conduct DIF analysis in different test situations to help test designers optimize or delete unfair items and provide better estimates of true abilities of test takers. The findings of this study call for an awareness movement in the area of language testing to enhance test bias understanding and to avoid the specific type of bias caused by extraneous variables besides the construct under investigation. By and large, test fairness explorations are needed to provide test designers and educators with reliable evaluations and fair judgments.

Finally, it is suggested that researchers employ a confirmatory DIF analysis besides an exploratory approach, seeking expert judgment to help formulate a hypothesis prior to undertaking DIF analysis. The researchers of this study recommend that using more than one type of statistical DIF detection method can verify the research findings, and preclude the situation wherein DIF remains undetected due to the limitations of a particular quantitative method. Furthermore, it is crucial to perform further research to determine whether these postulations can be verified with larger pools of items.

The results of the present study like any other studies endure some limitations as well, which can influence the findings and restrict the generalizability of the conclusions. Considering sampling procedure, further research could focus on generalization of the results by random selection of the sample of the study. This study made an effort to make a contribution to the DIF literature by providing information about DIF with an Iranian sample, and it is not clear whether DIF finding may be common across nationalities. Hence, other future research can be conducted to investigate DIF across gender and modes of learning among different nationalities.

Abbreviations

DIF	Differential item functioning
EFL	English as a Foreign Language
IRT	Item response theory
NUDIF	Non-uniform differential item functioning
PCA	Principal component analysis
UDIF	Uniform differential item functioning

Authors' contributions

EM conducted the main study. Her co-authors contributed to the writing and revision of the manuscript. The authors read and approved the final manuscript.

Authors' information

Elahe Moradi is a PhD candidate at Ferdowsi University of Mashhad, Iran. Her research interests include Language Testing, Psycholinguistics, and Sociolinguistics.

Dr. Zargham Ghabanchi holds a PhD in Applied Linguistics from the University of Liverpool, the UK. He has a chair at Ferdowsi University of Mashhad. He has published several articles.

Professor Reza Pishghadam is a professor of language education and a courtesy professor of educational psychology at Ferdowsi University of Mashhad, Iran. In 2010, he was classified as the distinguished researcher of humanities in Iran. In 2014, he also received the distinguished professor award from Ferdowsi Academic Foundation, Iran.

Funding
This study received no funding.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 25 April 2022 Accepted: 3 September 2022
Published online: 13 September 2022

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessment based on. *Language Assessment Quarterly*, *5*(2), 120–141.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB Listening Test. *Language Assessment Quarterly*, *8*(4), 361–385.
Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, *30*(2), 8–24. https://doi.org/10.1080/10904018.2015.1056876.
Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human Sciences*, (3rd ed.,). Mahwah: L. Erlbaum.
Bordbar, S. (2021). Gender differential item functioning (GDIF) analysis in Iran's university entrance exam. *English Language in Focus (ELIF)*, *3*(1), 49–68.
Breland, H., Lee, Y. W., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups (ETS Research Rep. No. 76)*. Princeton: Educational Testing Service.
Butler, Y. G., & Iino, M. (2021). Fairness in college entrance exams in Japan and the planned use of external tests in English. In *Challenges in language testing around the world*, (pp. 47–56). Singapore: Springer.
Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park: SAGE Publications.
Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. *ETS Research Report Series*, (2), i–143. https://doi.org/10.1002/j.2333-8504.1992 Tb01495.x.
Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*(2), 155–163.
Chubbuck, K., Curley, W. E., & King, T. C. (2016). Who's on first? Gender differences in performance on the SAT test on critical reading items with sports and science content. *ETS Research Report Series*, *16*(2), 1–116. https://doi.org/10.1002/ets2.12109.
Engelhard, G. (2012). *Invariant measurement: using Rasch models in the social, behavioral, and health sciences*. New York and London: Routledge.
Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148.
Jiao, H., & Chen, F. (2013). *Differential item and Testlet functioning analyses, the companion to language assessment, Wiley*.
Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation*, *19*(2), 158–169.
Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*, 23–48.
Lawrence, I. M., & Curley, W. E. (1989). *Differential item functioning for males and females on SAT-Verbal Reading sub score items: Follow-up study (ETS Research Report 89–22)*. Princeton: Educational Testing Service.
Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122.
Linacre, J. M. (2012). *A user guide to WINSTEPS MINISTEPS Rasch-model computer programs*. Chicago: Winsteps.com.
Linacre, J. M. (2018). *A user's guide to WINSTEPS*. Chicago: Winsteps.com.
Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415–437.

MacIntyre, P., Baker, S., Clement, R., & Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high s chool French immersion students. *Language Learning*, *53*(S1), 537–564. https://doi.org/10.1111/1467-9922.00194.
Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*, *46*(9), 3228–3237. https://doi.org/10.1016/j.measurement.2013.06.020.
Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology*, *35*(3), 299–314.
Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*(5), 793–817.
McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language Assessment*. Oxford: Oxford University Press.
McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (pp. 13–103). New York: ACE and Macmillan.
Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302.
Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*, 7. https://doi.org/10.1186/s40468-020-00105-2.
Oliveri, M. E., 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. https://doi.org/10.1080/08957347.2017.1391258.
Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson Test of English Academic. *Educational Research and Evaluation*, *18*(3), 211–229. https://doi.org/10.1080/13803611.2011.650921.
Pae, T.I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53-73.
Philips, D. (2005). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests*. White Plains: Longman.
Rashvand Semiyari, S., & Ahangari, S. (2022). Examining Differential Item Functioning (DIF) For Iranian EFL test takers with different fields of study. *Research in English Language Pedagogy*, *10*(1), 169–190. https://doi.org/10.30486/relp.2021.1935588.1295.
Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12–29.
Sehmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67–81.
van Langen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. *Comparative Education*, *41*(3), 329–350.
Wright, B. D., & Stone, M. H. (1988). *Identification of item bias using Rasch measurement. (Research Memorandum No. 55)*. Chicago: MESA Press.
Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wilmington: Wide Range.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465.
Zhu, X., & Aryadoust, V. (2019). Examining test fairness across gender in a computerized reading test: A comparison between the Rasch-based DIF technique and MIMIC. *Papers in Language Testing and Assessment*, *8*(2), 65–90.
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*, *35*(2), 412–436.
Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.
Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136–151. https://doi.org/10.1080/15434303.2014.972559.

Publisher's Note