Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 14, No. 1; 2024, March

ISSN 2476-5880
https://www.ijlt.ir

Applying IRT Model to Determine Gender and Discipline-based DIF and DDF: A Study of the IAU English Proficiency Test

Sarallah Jafaripour[1] , Omid Tabatabaei[2]*, Hadi Salehi[3] , Hossein Vahid Dastjerdi[4]

ARTICLE INFO

Article History:
Received: July 2023
Accepted: August 2023

KEYWORDS
Differential Distractor Functioning (DDF);
Differential Item Functioning (DIF);
English Proficiency Test (EPT);
Item Response Theory (IRT);
Test Bias

ABSTRACT

The purpose of this study was to examine gender and discipline-based Differential Item Functioning (DIF) and Differential Distractor Functioning (DDF) on the Islamic Azad University English Proficiency Test (IAUEPT). The study evaluated DIF and DDF across genders and disciplines using the Rasch model. To conduct DIF and DDF analysis, the examinees were divided into two groups: Humanities and Social Sciences (HSS) and Non-Humanities and Social Sciences (N-HSS). The results of the DIF analysis showed that four out of 100 items had DIF across gender, and two items had discipline DIF. Additionally, gender DDF analysis identified one item each for Options A, B, and C, and four items for Option D. Similarly, the discipline DDF analysis revealed one item for Option A, three items for Option B, four items for Option C, and three items for Option D. The findings of this study have significant implications for test developers. The identification of potential biases in high-stakes proficiency tests can help ensure fairness and equity for all examinees. Furthermore, identifying gender DIF can shed light on potential gender-based gaps in the curriculum, highlighting areas where male or female learners may be disadvantaged or underrepresented in terms of knowledge or skills.

1. Introduction

The present study was an attempt to provide a detailed account of Differential Item Functioning (DIF) as well as Differential Distractor Functioning (DDF) based on the gender and discipline of the examinees in order to ensure fairness in the Islamic Azad University English Proficiency Test (IAUEPT).

Test fairness which focuses on equal treatment and unbiased outcomes is considered an important aspect of test validity. Validity is one of the vital discussions in language testing which is becoming difficult to ignore. It is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores (Messick, 1989). Furthermore, test fairness is an extremely important issue

[1] English Department, Najafabad Branch, Islamic Azad University, Najafabad, Iran, Email: sarallah.jafaripour.1361@gmail.com
[2] English Department, Najafabad Branch, Islamic Azad University, Najafabad, Iran, Email: tabatabaeiomid@yahoo.com
[3] English Department, Najafabad Branch, Islamic Azad University, Najafabad, Iran, Email: hadisalehi1358@yahoo.com
[4] English Department, Najafabad Branch, Islamic Azad University, Najafabad, Iran, Email: h_vahid@yahoo.com

Cite this paper as: Jafaripour, S., Tabatabaei, O., Salehi, H., & Vahid Dastjerdi, H. (2024). Applying IRT model to determine gender and discipline-based DIF and DDF: A study of the IAU English proficiency test. *International Journal of Language Testing*, *14*(1), 56–74. https://doi.org/10.22034/IJLT.2023.407117.1268

in language testing which is largely related to test validation (Amirian et al., 2014; Perrone, 2006). DIF, a key element to evaluate the fairness and validity of educational and psychological tests, happens when two groups of equal ability levels are not equally able to correctly answer an item (Karami, 2012).

DDF, which is a method to examine the examinees' responses for group differences in distractor selection rates (Green et al., 1989), sheds light on the reasons for DIF and provides evidence that the cause of DIF has roots in the features of accurate answers (Jamalzadeh et al., 2021).

DIF only indicates that there is a difference in item performance between groups but does not provide a definitive explanation for the observed differences. In fact, other factors such as cultural or linguistic differences may also contribute to the observed DIF (Camilli & Shepard, 1994). As declared by Holland and Thayer (1988), further investigations like qualitative research or expert reviews are essential to understand the sufficient reasons of DIF and determine if bias exists. Regarding DDF, a number of studies have acknowledged the fact that relying solely on DDF to determine item bias brings about some limitations as well (Deng, 2020; Koon & Kamata, 2013; Middleton & Laitusis, 2007).

Item Response Theory (IRT) is one of the models for DIF and DDF detection that has recently been widely applied (Bakytbekovich et al., 2023; Batty, 2015; Fayers & Machin, 2007). The chief causes for this large-scale employment are attributed to its ability to account for item characteristics, flexibility in modeling DIF and DDF, and capacity for offering a unified framework for modeling item responses (Bortolotti et al., 2013; Hambleton & Swaminathan, 2013). However, it is important to consider the limitations of the IRT model, such as assumptions regarding model fit and the need for large sample sizes (Singh, 2004).

Although the vast majority of the work in this area has focused on gender DIF (e.g., Huang et al., 2022; Metu, 2020; Ravand et al., 2019), there is limited research investigating both DIF and DDF (e.g., Sandersfeld, 2020). As a matter of fact, scant attention has been paid to DIF and DDF detection in high-stakes proficiency tests administered to male and female test takers with different fields of study. To bridge the chasm, the current study employed the IRT model to find out whether IAUEPT, a high-stakes proficiency test administered to PhD candidates, reveals significant DIF and DDF in favor of a particular gender or discipline group. Accordingly, the present study addresses the following questions:

Q1. Does participants' gender cause significant DIF in IAUEPT?
Q2. Does participants' discipline cause significant DIF in IAUEPT?
Q3. Does participants' gender cause significant DDF in IAUEPT?
Q4. Does participants' discipline cause significant DDF in IAUEPT?

2. Review of Literature

DIF analysis has generally been accepted as a standardized way of validating the tests in the broader field of assessment (Shahmirzadi, 2023). As declared by Salehi and Tayebi (2012), DIF procedures are considered the new dominant psychometric methods to address fairness in standardized, achievement, aptitude, and license testing. Numerous scholars have stressed the relationship between DIF and DDF, both of which refer to the analysis of test items in educational assessments. Over the past few years, a number of studies have been carried out in an attempt to assess the impact of distractors (e.g., Adibatmaz & Yildiz, 2020; Baghaei & Dourakhshan, 2016) and analyze DIF for group differences (e.g., Amirian, 2020; Giguère et al., 2023). However, studies on DIF and DDF, some of which are reviewed here, have yielded contradictory findings.

2.1. Previous Studies on DIF

Amirian et al. (2014) attempted to examine whether the University of Tehran English Proficiency Test (UTEPT) indicated significant gender-related DIF. They found that the effect size of DIF was mostly insignificant to render the test unfair. This finding is not congruent with the work of Barati and Ahmadi (2010) who investigated DIF on the Special English Test of the Iranian National University Entrance Exam (INUEE) utilizing a one-parameter IRT model. The effect of gender and subject area was taken into account, and the findings confirmed the presence of DIF on this test.

In recent years, researchers have become increasingly interested in the assessment of DIF, considering the examinees' fields of study. Amidst numerous studies on the discipline DIF, Estaji and Zhaleh (2020) performed a study to examine the impact of the test takes' fields of study on the reading

section of the English subtest of the Iranian University Entrance Exam (IUEE) for MA in English majors. To meet the objectives of the study, one and two-parameter logistic IRT models were applied to examine DIF. The one-parameter DIF analysis results showed that out of 20 items in the reading section, only three items displayed DIF toward the examinees based on their disciplines. Nevertheless, two-parameter DIF analysis results indicated that all the items of the reading section presented DIF toward the test takers. Similarly, Rashvand Semiyari and Ahangari (2022) in their study on discipline DIF found that Science examinees outperformed the Humanities. Additionally, they inferred that the exam was statistically easier for Science test-takers at 0.05 level.

2.2. Previous Studies on DDF

Bakytbekovich et al. (2023) conducted a study to analyze items' distractors of a grammar test applying the Rasch model. The results demonstrated an acceptable fit to the Rasch model and high reliability. Moreover, malfunctioning distractors were identified. Their findings are consistent with the study run by Sandersfeld (2020) who examined items from Grades 3, 6, and 9 of the 2018-2019 administration of the Iowa Statewide Assessment of Student Progress (ISASP) mathematics test. Each item's distractors were examined for DDF between test takers of the two test delivery modes within the total population and the selected demographic subgroups. It was found that a total of eleven items displayed evidence of DDF between modes for at least one distractor within the total population, and a total of thirty-one items indicated evidence of DDF between modes for at least one distractor within at least one demographic group.

To summarize, the above-mentioned studies were reviewed in light of DIF and DDF. Compared to DIF, DDF does not involve a great deal of literature. For example, a few studies on recommended sample sizes for DDF analysis were found at the time of accomplishment of this article. According to Sandersfeld (2020), if an item includes a special distractor pulling the focal group toward it often enough to cause DIF detection, then understanding what aspects these distractors have that pull the focal group away can greatly benefit test developers working to produce fair testing instruments. As stated by Penfield (2010), the existence of DIF in a multiple-choice item displays a violation of invariance, but the results of a DIF analysis alone do not provide enough data as to where among the response choices the DIF impact is being manifested. So far, a limited number of studies have examined the combination of DIF and DDF in analyzing English proficiency tests. Through considering a number of male and female PhD candidates from different disciplines who sit for IAUEPT, the standardization of this test should receive top priority in language assessment investigation. Hence, the examination and improvement of this test can undoubtedly be considered a worthwhile research outcome.

3. Method

3.1. Participants

The participants of the present study who took the spring 2021-2022 version of IAUEPT were 1069 PhD candidates of Humanities (n= 176), Medicine (n= 155), Physical Education (n= 207), Architecture (n= 151), Educational Sciences (n= 210), and Agriculture (n= 170) disciplines with an age range of 23 and 49. Among the participants, 684 examinees were females and 385 were males. The IAUEPT is designed to measure the general ability of various test takers in different disciplines in order to select the best exit program for PhD candidates.

3.2. Instrumentation

Islamic Azad University English Proficiency Test (IAUEPT), the results of which are applied for making high-stakes decisions about PhD candidates, requires precise validation procedures as well as objective planning and design because such results will have profound educational and ethical consequences. This test consists of 100 multiple-choice items in four areas of *Vocabulary* (25 items), *Structure Part I* (25 items), *Structure Part II* (15 items), and *Reading Comprehension* (35 items). IAUEPT was analyzed with the SPSS program version 22, and the mean of the population (n= 1069) for the whole test was 78.33 out of 100. The Cronbach's alpha reliability of the whole test was .84 which was quite high. The descriptive statistics and reliabilities for the test sections can be seen in Table 1.

Table 1
Reliability Estimates for the IAUEPT

Test	No. Items	Mean	SD	Variance	Min.	Max.	Range	Reliability
Vocabulary	25	21.6	2.1	4.41	15	25	10	.44
Structure I	25	19.6	3.1	9.61	1	25	24	.68
Structure II	15	8.6	2	4	2	14	12	.26
Reading	35	28.5	6.6	43.56	0	35	35	.92
Total	100	78.33	8.88	78.87	45	94	49	.84

For the purpose of this study, DIF and DDF across gender and discipline were evaluated by applying the Rasch model (Rasch, 1960/1980). Moreover, the Winsteps computer program (Linacre, 2009) was applied to estimate the model.

3.3. Procedures

Firstly, the anonymous IAUEPT answer sheets of the targeted examinees were received from the examination office in IAU in Tehran. The permission to get access to answer sheets was gained through corresponding with the authorities of Islamic Azad University in the central department and undertaking to keep the data of PhD candidates private. Then 1069 male and female test takers were divided into two discipline groups because DIF should be conducted across two groups. Accordingly, Humanities, Physical Education, and Educational Sciences were combined into one group labeled Humanities and Social Sciences (HSS), and Medicine, Architecture, and Agriculture were combined into a group labeled Non-Humanities and Social Sciences (N-HSS). Next, DDF was evaluated for Options A, B, C, and D across gender and discipline separately in three steps:

1. The test was rescored as 'option a=1', 'other options=0'.
2. Standard DIF across a grouping variable (e.g., gender) was run.
3. Items demonstrating DIF were identified. If an item indicates DIF say, in favor of females, this means that option 'a' has attracted more females than males. If option 'a' happens to be the correct choice, DDF and DIF coincide in this item and DDF reduces to DIF.

The procedure above was repeated for the other options across both grouping variables. All in all, the data was analyzed through the Rasch model for the presence of gender and discipline DIF and DDF respectively.

4. Results

4.1. Item Measure and Fit Values

The following is the table of item statistics and fit values. The interpretation follows the table.

Table 2
Item Measure and Fit Values

Item	Measure	SE	Infit MNSQ	Outfit MNSQ	PT-Measure CORR.
1	-.47	.22	1.14	1.24	.02
2	-1.23	.29	1.03	.93	.16
3	.51	.17	1.01	.96	.29
4	-.38	.21	1.09	1.04	.12
5	-.68	.24	1.04	.99	.18
6	.05	.19	1.09	1.09	.13

7	-.80	.25	1.07	1.05	.12
8	.31	.17	1.00	.96	.29
9	.65	.16	1.03	1.03	.23
10	-5.10	1.82	MINIMUM MEASURE		.00
11	-2.78	.58	.96	.40	.24
12	-.29	.21	1.07	.96	.18
13	.02	.19	1.05	1.14	.17
14	-2.06	.42	.97	.63	.23
15	-.80	.25	1.14	1.41	-.04
16	-1.07	.27	1.05	1.09	.11
17	-.52	.22	1.07	1.18	.11
18	1.13	.15	1.02	1.03	.25
19	-2.06	.42	1.02	1.19	.07
20	-.43	.22	1.07	1.03	.15
21	-1.32	.30	1.08	1.26	.02
22	-1.07	.27	.99	.85	.24
23	-1.23	.29	1.09	1.35	.01
24	2.07	.15	1.16	1.23	.02
25	-5.10	1.82	MINIMUM MEASURE		.00
26	.46	.17	1.14	1.15	.07
27	2.45	.15	.99	1.08	.25
28	-1.32	.30	1.00	.91	.20
29	-.13	.20	1.09	1.04	.15
30	.31	.17	1.04	1.03	.22
31	-1.41	.32	.97	.76	.25
32	-.25	.20	1.01	.88	.27
33	-2.78	.58	.97	.42	.23
34	-.02	.19	1.07	1.03	.18
35	.75	.16	1.10	1.09	.15
36	2.33	.15	1.02	1.09	.21
37	-.74	.24	1.03	1.21	.14
38	-.57	.23	1.07	1.13	.11
39	1.88	.15	1.09	1.16	.12
40	-.05	.19	1.09	1.07	.13
41	-.57	.23	1.05	1.10	.14
42	-1.63	.35	.97	.77	.24
43	-.43	.22	1.10	1.23	.06
44	-.05	.19	1.11	1.12	.10
45	2.67	.16	.96	1.04	.29
46	-1.90	.39	.99	.84	.18
47	-.05	.19	1.06	1.05	.17
48	-.34	.21	1.00	.85	.28
49	.09	.18	1.07	1.02	.18
50	-.74	.24	1.05	.99	.16
51	.09	.18	1.13	1.17	.07
52	.54	.17	1.13	1.11	.11
53	1.83	.15	1.03	1.09	.22
54	2.43	.15	1.16	1.35	-.02
55	3.00	.17	1.05	1.16	.13
56	.40	.17	1.05	1.01	.22
57	1.41	.15	1.09	1.12	.14
58	.92	.15	1.03	1.01	.25
59	2.27	.15	.99	.99	.29

60	1.81	.15	1.02	1.07	.24
61	-.21	.20	1.13	1.33	.02
62	.09	.18	1.05	.99	.21
63	2.83	.16	1.12	1.36	.01
64	.75	.16	1.10	1.17	.11
65	1.52	.15	1.02	.99	.27
66	-1.00	.27	.88	.84	.36
67	.05	.19	.95	1.07	.29
68	1.71	.15	1.04	1.04	.23
69	.05	.19	.95	.97	.32
70	.78	.16	.97	.99	.32
71	.75	.16	.97	.98	.32
72	-1.15	.28	.78	.39	.57
73	.97	.15	.95	.92	.37
74	-.57	.23	.91	.81	.36
75	-1.90	.39	.81	.23	.52
76	.43	.17	.96	.95	.34
77	.22	.18	.87	.80	.47
78	.85	.16	.88	.83	.48
79	.12	.18	.92	.94	.36
80	.46	.17	.97	.97	.32
81	.43	.17	.96	.97	.33
82	-.68	.24	.82	.67	.50
83	-1.15	.28	.86	.64	.42
84	-.34	.21	.83	.76	.48
85	-.09	.19	.92	.92	.36
86	1.94	.15	1.00	.97	.29
87	.28	.18	.90	.90	.41
88	-1.23	.29	.81	.56	.48
89	-1.41	.32	.78	.31	.57
90	-.74	.24	.79	.53	.56
91	-.63	.23	.79	.57	.56
92	-.86	.25	.85	.64	.45
93	.49	.17	.96	.97	.33
94	-.05	.19	.89	.89	.40
95	-.93	.26	.76	.42	.61
96	.78	.16	.97	.98	.32
97	-1.76	.37	.80	.25	.54
98	-.13	.20	.87	.81	.44
99	-1.07	.27	.75	.38	.62
100	.19	.18	.83	.81	.51

The column ‘measure’ displays the item difficulty. The higher the measure, the more difficult the items is. Based on Table 2, Item 55 is the most difficult item and Item 25 is the easiest item. The column ‘S.E.’ indicates the standard error of the item measures. The smaller the S.E., the more accurately the item difficulties have been estimated. Infit and outfit mean square values show the extent to which the items fit the Rasch model (Baghaei et al., 2017). Values smaller than 1.30 are acceptable. As Table 2 shows, Items 15, 54, 61, and 63 have outfit mean square values greater than 1.30 and do not fit the Rasch model. The column ‘PT-Measure CORR.’ indicates the point-biserial correlation between the items and person ability measures. The higher point-measure correlations, the better discriminating the item is. Values greater than .20 are acceptable (Linacre, 2023). Based on point-measure correlations, many items in the test have poor quality and must be discarded.

4.2. Addressing Research Question One

The first question was posed to figure out if participants' gender causes significant DIF in IAUEPT. Table 3 displays the relevant statistics for DIF across gender. The column 'DIF Measure' on the left indicates the difficulty of each item for females, and 'DIF Measure' towards the right shows the item difficulty for males. 'Contrast' indicates the difference in two difficulty measures. This difference is tested for statistical significance with a t-test. The column 't' displays the value of *t* statistic, column 'd.f.' shows the degrees of freedom, and column 'Prob.' shows the p-value. Note that, because of space restrictions, only items revealing gender DIF appear in Table 3.

Table 3
DIF Statistics Across Gender

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I3	Female	.12	Male	.83	-.70	-2.07	202	.0396
I6	Female	-.44	Male	.42	-.86	-2.20	199	.0286
I60	Female	1.30	Male	2.30	-1.01	-3.37	204	.0009
I66	Female	-.54	Male	-1.73	1.20	1.97	196	.0499

The null hypothesis states that participants' gender does not cause significant DIF in IAUEPT. However, the alternative hypothesis states that participants' gender brings about significant DIF in IAUEPT. The results of Table 3 indicate that the participants' gender caused significant DIF in IAUEPT since p-values smaller than .05 display that the difficulty difference between males and females for an item is significant and the item had DIF. The column 'Prob.' in Table 3 shows that Items 3, 6, 60, and 66 have DIF across gender. Items 3, 6, and 60 exhibit DIF in favor of females but Item 66 is in favor of males. The smaller the DIF measure, the easier the item is. Therefore, participants' gender caused significant DIF in IAUEPT, and the null hypothesis was rejected.

4.2.1. Graphical Displays of DIF Across Gender

Figure 1
Male and Female ICCs for Item 3

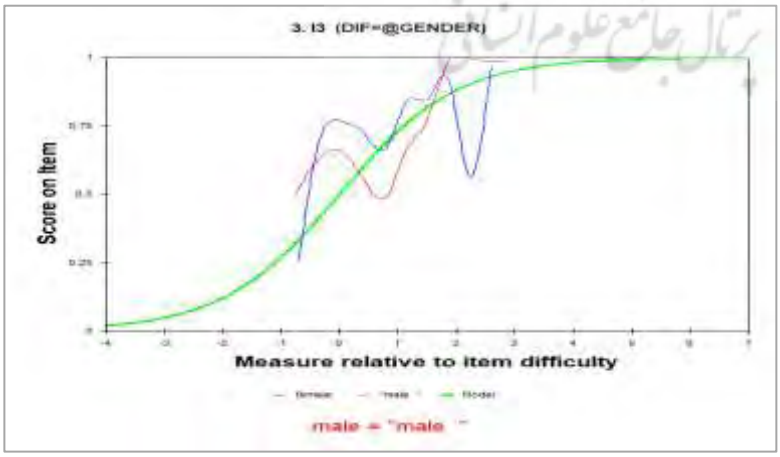


Figure 1 depicts the Item Characteristic Curves (ICC) for Item 3 across males and females. The blue curve represents the ICC for females and the red curve represents the ICC for males. The horizontal axis displays person ability, and the vertical axis shows the probability of success or the probability of a correct response. The monotonicity assumption of IRT mandates that as the ability increases, the probability of a correct response should increase too (Baghaei, 2021). In fact, an ideal item should behave like the S-shaped green ICC, which indicates the ideal Rasch model required ICC. For an item not to have DIF the male and female ICCs should overlap. As the figure indicates the ICCs for males and females diverge at certain points.

4.3. Addressing Research Question Two

The second research question was posed to investigate whether participants' discipline causes significant DIF in IAUEPT. The results are shown in the following Table. For the sake of convenience, only items showing discipline DIF are indicated in the following Table.

Table 4
DIF Statistics Across Discipline

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I76	HSS	.91	N-HSS	-.15	1.06	3.00	204	.0030
I80	HSS	.81	N-HSS	.07	.75	2.18	204	.0305

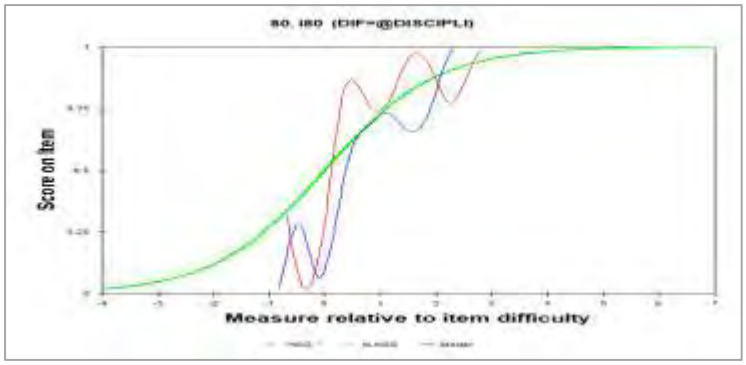
Table 4 indicates the relevant statistics for DIF across disciplines. The column 'DIF Measure' on the left displays the difficulty of each item for the HSS group, and 'DIF Measure' towards the right indicates item difficulty for the N-HSS.

The null hypothesis expresses that participants' discipline does not cause significant DIF in IAUEPT. In contrast, the alternative hypothesis puts into words that participants' discipline leads to significant DIF in IAUEPT. The results of Table 4 indicate that the participants' discipline caused significant DIF in IAUEPT since p-values smaller than .05 show that the difficulty difference between HSS and N-HSS groups for an item is significant and the item has DIF across disciplines. The column 'Prob.' in Table 4 demonstrates that Items 76 and 80 have discipline DIF. Both items have DIF in favor of N-HSS group. The smaller the DIF measure, the easier the item is. Consequently, participants' discipline caused significant DIF in IAUEPT, and the null hypothesis was rejected.

4.3.1. Graphical Displays of DIF Across Discipline

Figure 2 depicts the ICCs for Item 80 across HSS and N-HSS groups. The blue curve represents the ICC for the HSS group, and the red curve stands for the ICC for the N-HSS group. The ICC needed for the ideal Rasch model is displayed in the S-shaped green ICC. For an item not to have DIF, the HSS, and the N-HSS ICCs should overlap. As the figure shows the ICCs for HSS and N-HSS diverge at certain points.

Figure 2
HSS and N-HSS ICCs for Item 80



4.4. Addressing Research Question Three

The third research question was put forward to examine if test takers' gender causes significant DDF in IAUEPT. For this purpose, DDF was evaluated for each option separately. DDF was run in three steps presented in the procedure section.

4.4.1. DDF Across Gender for Option A

Table 5 indicates the DDF results for Option A across gender. 'DIF Measure' on the left reveals the degree to which females have selected 'Option A' while 'DIF Measure' towards the right shows the degree to which males have selected 'Option A'. The smaller the DIF measure, the more popular the option has been in the group. Due to limited space, Table 5 solely displays the item suffering from DDF across gender.

Table 5
DDF for Option A Across Gender

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I85	Female	-4.23	Male	-3.27	-.96	-2.21	188	.0286

The Prob. column displays that only Item 85 has a p-value smaller than .05. This suggests that Option A in Item 85 has DDF across gender. This item has a DIF measure of -4.23 for females and -3.27 for males. This indicates that Option A in Item 85 has been more popular among females. The difference in the DIF measure across the two groups (DIF Contrast=-.96) is statistically significant (t=-2.21, df=188, p=.02).

4.4.2. DDF Across Gender for Option B

Table 6 presents the DDF results for Option B across gender. The Prob. column points out that only Item 60 has a p-value smaller than .05. This suggests that Option B in Item 60 has DDF across gender. This item has a DIF measure of -2.09 for females and -1.16 for males. This indicates that Option B in Item 60 has been more popular among females. The difference in the DIF measure across the two groups (DIF Contrast=-.93) is statistically significant (t=-3.21, df=203, p=.001). Only Item 60 is included in the following Table due to space restrictions.

Table 6
DDF for Option B Across Gender

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I60	Female	-2.09	Male	-1.16	-.93	-3.21	203	.0016

4.4.3. DDF Across Gender for Option C

Table 7 demonstrates the DDF results for Option C across gender. The Prob. column uncovers that Item 85 has a p-value smaller than .05. This proposes that Option C in Item 85 has DDF across gender. This item has a DIF measure of 1.42 for females and .19 for males. This indicates that Option C in this item has been more popular among males. The difference in the DIF measure across the two groups (DIF Contrast=1.23) is statistically significant (t=2.46, df=183, p=.01).

Table 7
DDF for Option C Across Gender

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I85	Female	1.42	Male	.19	1.23	2.46	183	.0148

4.4.4. DDF Across Gender for Option D

Table 8 makes the DDF results visible for Option D across gender. The Prob. column reveals that Items 3, 9, 54, and 60 have p-values smaller than .05. This suggests that Option D in these four items has DDF across gender. Note that, due to the limited space, Table 8 contains only the items that suffer from DDF across gender.

Table 8
DDF for Option D Across Gender

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I3	Female	-3.00	Male	-2.34	-.67	-2.00	201	.0473
I9	Female	-.59	Male	.38	-.97	-2.76	203	.0063
I54	Female	1.53	Male	.30	1.23	2.46	193	.0149
I60	Female	-.39	Male	-1.20	.81	2.64	202	.0090

Item 3 has a DIF measure of -3 for females and -2.34 for males. This indicates that Option D in Item 3 was more popular among females. The difference in the DIF measure across the two groups (DIF Contrast=-.67) is statistically significant (t=-2, df=201, p=.04). Item 9 has a DIF measure of -.59 for females and .38 for males. This indicates that Option D in Item 9 was more popular among females.

The difference in the DIF measure across the two groups (DIF Contrast=-.97) is statistically significant (*t*=-2.76, *df*=203, *p*=.006). Item 54 has a DIF measure of 1.53 for females and .30 for males. This indicates that Option D in Item 54 was more popular among males. The difference in the DIF measure across the two groups (DIF Contrast=1.23) is statistically significant (*t*=2.46, *df*=193, *p*=.01). Item 60 has a DIF measure of -.39 for females and -1.20 for males. This indicates that Option D in Item 60 was more popular among males. The difference in the DIF measure across the two groups (DIF Contrast=.81) is statistically significant (*t*=2.64, *df*=202, *p*=.009).

The null hypothesis proclaims that participants' gender does not cause significant DDF in IAUEPT. On the contrary, the alternative hypothesis asserts that participants' gender brings about significant DDF in IAUEPT. It is obvious from the results of Tables 5, 6, 7, and 8 that the above-mentioned items suffer from DDF across gender. Consequently, participants' gender caused significant DDF in IAUEPT, and the null hypothesis was rejected.

4.5. Addressing Research Question Four

The fourth research question was posed to figure out whether participants' discipline causes significant DDF in IAUEPT. To achieve this aim, DDF was evaluated for each option separately. Accordingly, the same procedure mentioned for addressing research question three was repeated for each option across disciplines. Notice that, like the previous tables, only items suffering from DDF across disciplines appear in the following Tables.

4.5.1. DDF Across Discipline for Option A

As explained earlier, discipline DDF was evaluated across two major discipline groups, namely, HSS and N-HSS. Table 9 presents the DDF results for Option A across discipline. 'DIF Measure' on the left indicates the degree to which the HSS group has selected 'Option A' while 'DIF Measure' towards the right shows the degree to which the N-HSS group has selected 'Option A'. The smaller the DIF measure, the more popular the option has been in a group.

Table 9
DDF for Option a Across Discipline

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I79	HSS	.53	N-HSS	2.22	-1.69	-2.52	187	.0127

The Prob. column in Table 9 demonstrates that only Item 79 has a p-value smaller than .05. This suggests that Option A in Item 79 has DDF across disciplines. This item has a DIF measure of .53 for HSS and 2.22 for the N-HSS group. This indicates that Option A in Item 79 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-1.69) is statistically significant (*t*=-2.52, *df*=187, *p*=.01).

4.5.2. DDF Across Discipline for Option B

Table 10 shows the DDF results for Option B across disciplines. The Prob. column in Table 10 reveals that Items 2, 44, and 76 have p-values smaller than .05. This suggests that Option B in these items has DDF across disciplines. Item 2 has a DIF measure of 1.25 for HSS and 3.66 for N-HSS group. This indicates that Option B in Item 2 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-2.42) is statistically significant (*t*=-2.28, *df*=181, *p*=.02). Item 44 has a DIF measure of 2.53 for HSS and .36 for N-HSS group. This indicates that Option B in Item 44 has been more popular among the N-HSS group. The difference in the DIF measure across the two groups (DIF Contrast=2.16) is statistically significant (*t*=3.12, *df*=163, *p*=.002). Item 76 has a DIF measure of -2.54 for HSS and -3.73 for N-HSS group. This indicates that Option B in Item 76 has

been more popular among the N-HSS group. The difference in the DIF measure across the two groups (DIF Contrast=1.19) is statistically significant (*t*=3.28, *df*=196, *p*=.001).

Table 10
DDF for Option B Across Discipline

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I2	HSS	1.25	N-HSS	3.66<	-2.42	-2.28	181	.0239
I44	HSS	2.53	N-HSS	.36	2.16	3.12	163	.0022
I76	HSS	-2.54	N-HSS	-3.73	1.19	3.28	196	.0012

4.5.3. DDF Across Discipline for Option C

Table 11 exhibits the DDF results for Option C across disciplines. The Prob. column in Table 11 reveals that Items 3, 48, 76, and 80 have p-values smaller than .05. This suggests that Option C in these items has DDF across disciplines. Item 3 has a DIF measure of 2.93 for HSS and 1.03 for N-HSS group. This indicates that Option C in Item 3 has been more popular among the N-HSS group. The difference in the DIF measure across the two groups (DIF Contrast=1.90) is statistically significant (*t*=2.39, *df*=175, *p*=.01). Item 48 has a DIF measure of .23 for HSS and 1.17 for N-HSS group. This indicates that Option C in Item 48 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-.94) is statistically significant (*t*=-2.01, *df*=201, *p*=.04). Item 76 has a DIF measure of -.58 for HSS and .54 for N-HSS group. This indicates that Option C in Item 76 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-1.11) is statistically significant (*t*=-2.98, *df*=196, *p*=.003). Item 80 has a DIF measure of -.41 for HSS and .64 for N-HSS group. This indicates that Option C in Item 80 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-1.05) is statistically significant (*t*=-2.70, *df*=196, *p*=.007).

Table 11
DDF for Option C Across Discipline

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I3	HSS	2.93	N-HSS	1.03	1.90	2.39	175	.0180
I48	HSS	.23	N-HSS	1.17	-.94	-2.01	201	.0461
I76	HSS	-.58	N-HSS	.54	-1.11	-2.98	196	.0033
I80	HSS	-.41	N-HSS	.64	-1.05	-2.70	196	.0075

4.5.4. DDF Across Discipline for Option D

Table 12 shows the DDF results for Option D across disciplines. The Prob. column in Table 12 displays that Items 26, 44, and 80 have p-values smaller than .05. This suggests that Option D in these items has DDF across disciplines. Item 26 has a DIF measure of .11 for HSS and 1.10 for N-HSS group. This indicates that Option D in Item 26 has been more popular among the HSS group. The difference

in the DIF measure across the two groups (DIF Contrast=-.99) is statistically significant (*t*=-2.22, *df*=200, *p*=.02). Item 44 has a DIF measure of -3.97 for HSS and -3.07 for N-HSS group. This indicates that Option D in Item 44 has been more popular among the HSS group. The difference in the DIF measure across the two groups (DIF Contrast=-.89) is statistically significant (*t*=-2.03, *df*=185, *p*=.04). Item 80 has a DIF measure of -2.41 for HSS and -3.26 for N-HSS group. This indicates that Option D in Item 80 has been more popular among the N-HSS group. The difference in the DIF measure across the two groups (DIF Contrast=.85) is statistically significant (*t*=2.40, *df*=197, *p*=.01).

Table 12
DDF for Option D Across Discipline

Item Number	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
I26	HSS	.11	N-HSS	1.10	-.99	-2.22	200	.0273
I44	HSS	-3.97	N-HSS	-3.07	-.89	-2.03	185	.0442
I80	HSS	-2.41	N-HSS	-3.26	.85	2.40	197	.0172

The null hypothesis declares that participants' discipline does not cause significant DDF in IAUEPT. However, the alternative hypothesis states that participants' discipline brings about significant DDF in IAUEPT. It is evident from the results of Tables 9, 10, 11, and 12 that the items explained in these tables suffer from DDF across disciplines. Therefore, participants' discipline caused significant DDF in IAUEPT, and the null hypothesis was rejected.

5. Discussion

The results of DIF analysis across gender indicated that Item 3 showed DIF in favor of females. There have been numerous studies conducted on the relationship between examinees' gender and their different performances on tests. Some studies have been aimed at the relationship between gender DIF and item content (Bolger & Kellaghan, 1990; Lane et al., 1996; Willingham & Cole, 1997). Item 3 came from the vocabulary section.

Item 3. A(n) ……. referee does not favor one team in relation to another.
A) intimate B) biased
C) subjective D) impartial

Item 3, which is a decontextualized language item, could not be justified by the notion of "*topic familiarity*". This item was in favor of females because research has revealed that females tend to perform better on tests that involve verbal ability and vocabulary skills (Hedges & Nowell, 1995). This advantage is believed to be due to differences in brain structure and function between males and females as well as differences in socialization and educational experiences (Halpern, 2013). Moreover, it seems that the reason why Item 3 advantaged females is because of the words "*favor*" and "*relation"*. A set of studies have certified that females outperform males on items that are of particular relevance to human relations and rights (e.g., Donlon, 1973; Mazzeo et al., 1993).

Item 60. A) We visited our relatives during our vacation.
B) Do you know where can I find the Public Library?
C) At first I thought I could never learn English.
D) The quickest way to travel is by air.

Item 60, which belongs to the structure section, has been designed in order to measure the examinees' grammatical knowledge of noun clauses. One possible explanation for why this item was in favor of female test takers is that it may have tapped into a language proficiency or sensitivity that is more commonly associated with females. Research has indicated that females tend to outperform males on tests of language ability, including tests of grammar, vocabulary, and reading comprehension (e.g., Hyde & Linn, 1988). This may be due to a variety of factors, including differences in brain structure and function, socialization processes, and educational experiences.

Another possible explanation is that Item 60 may have been biased in some way that favored females. DIF can be caused by a variety of factors, including differences in cultural background, language proficiency, and test-taking strategies. In the case of Item 60, it is possible that the incorrect options were worded in a way that was more confusing or misleading for male test takers, or that the correct option relied on a subtle grammatical or syntactical rule that was more familiar to female test takers. Apparently, this item was more popular among females because of the phrase "*public library"* in Option B, which is the correct one. According to Amirian et al. (2014), such words and phrases reveal social interactions which is a topic of women's interest.

5.1. Rationales Behind Different Test Performances Among Males and Females

Numerous studies have displayed that females tend to have better verbal abilities, while males tend to have better spatial abilities (e.g., Halpern, 2013). This may lead to differences in the way that males and females process and understand language, which could affect their performance on language tests. A feasible interpretation is that there may be cultural or societal factors that influence the language skills of males and females. For example, a number of scholars have noted that girls tend to receive more encouragement and support for language learning than boys do (e.g., Oga-Baldwin & Fryer, 2020). This may lead to differences in the amount of exposure and practice that males and females have with language, which could affect their performance on language tests.

The role of brain structure and cognition in different language test performances among male and female examinees has been a topic of interest for researchers for many decades. One theory suggests that men and women have different brain structures that affect their language abilities. Shaywitz et al. (1995) pointed out that women have a larger language processing area in the brain than men, which could explain why women tend to perform better on language tests. Some researchers argued that the differences in language performance between men and women may be due to social and cultural factors, rather than biological factors. According to Maccoby and Jacklin (1974), differences in language performance between boys and girls were largely due to differences in socialization.

In sum, the reasons for gender-based differences in language test performance are likely to be complicated and multifaceted. While there is evidence to suggest that biological factors such as brain structure and cognition play a role in language test performance, it is important to consider the influence of social and cultural factors as well.

Item 26. Amir …….…… to Islamic Azad University where he studied history.

A) went	B)goes
C) will be gone	D)was going

The results for Option D across discipline DDF showed that Option D in Item 26 has been more accepted among the HSS group. This item whose goal was to assess the knowledge of the examinees concerning different types of tenses came from the structure section. Humanities and social sciences students often study historical events and narratives which require them to apply past tense verbs. Accordingly, they may be more familiar with past continuous tense (e.g., "*was going*") than other tenses. Additionally, numerous investigations have acknowledged the fact that past continuous tense is commonly applied in historical events, literature, and narratives. The participants of HSS group may have been exposed to this usage of the past continuous tense in their studies, making it a more natural choice for them. For example, a study run by Berman and Slobin (2013) found that the past continuous tense was applied more frequently in narratives and historical events.

Item 76. The most suitable title for this passage is ……. .
A) The Magic of Recycling
B) Methods of Waste Management: Pros and Cons
C) Recycling, Landfilling, or Composting: The Choice is Yours
D) How to Save the Earth by Recycling and Composting

The results for Option C across discipline DDF indicated that Option C in Item 76 was more popular among the HSS group. This item belonged to the reading passage whose title is "*Methods of Waste Management: Pros and Cons*". It is supposed that the HSS group was attracted to the words "*recycling*", "*landfilling*", and "*composting*" in Option C because these concepts involve human behavior, attitudes, and values toward the environment. Moreover, they involve social and economic factors such as government policies, consumer behavior, and waste management practices. As declared by Gregson and Crang (2010), social science and humanities perspectives can be applied in order to pose questions about how waste comes into being via relationships, language, politics, practices, and structures.

5.2. Rationales Behind Different Test Performances Among Students with Different Fields of Study

There are several theories and rationales for why students with different disciplines may have different performances on language tests. One of the reasons is related to the amount of exposure to the language. Students who study majors that require more language proficiency, such as literature or linguistics, may have more exposure to the language in their coursework and may therefore perform better on language tests. On the other hand, students in fields that require less language proficiency, such as science or engineering, may have less exposure to the language and may struggle more with language tests. Secondly, cognitive abilities such as working memory, attention, and processing speed can influence language proficiency (Daneman & Merikle, 1996). Students in disciplines that need more cognitive abilities, such as philosophy or mathematics, may have an advantage on language tests due to their stronger cognitive skills. The next reason is related to the fact that students who are more motivated to learn a language may perform better on language tests. For instance, students in fields that require more international communication, such as business or diplomacy, may be more motivated to learn a language and may therefore perform better on language tests. Finally, different disciplines may require different learning strategies, which can affect students' performance on language tests. For instance, a student studying law may need to focus on reading and analyzing legal texts, while a student studying business may need to focus on oral communication and negotiation skills.

6. Conclusion

The goal of this study was to test standardization through examining DIF on IAUEPT based on the gender and discipline of candidates. Moreover, DDF was analyzed to clarify the reasons of DIF. Firstly, DIF analysis identified four out of 100 items exhibiting DIF across gender in the analysis of all data. Indeed, while three items favored females, just one item was in favor of male test takers. Secondly, DIF statistics across disciplines determined that two items, both of which were in favor of N-HSS group, had discipline DIF. Next, DDF analysis was run to analyze gender and discipline DDF for each option separately. DDF analysis across gender identified one item for Options A, B, and C, whereas four items for Option D were recognized. Among the options of these items, four were more accepted among females, and three were more popular among males. Finally, the results for DDF analysis across disciplines identified one item for Option A, three items for Option B, four items for Option C, and three items for Option D. Among the options of these items, seven options were more popular among the HSS group, while four options were more accepted among the N-HSS group.

Regarding the test sections, the results of gender DIF revealed that females were more favored than males in the items of structure and vocabulary sections, while in the items of reading comprehension part, males were more favored. Then the findings of discipline DIF indicated that the N-HSS group was more favored than the HSS group in the items of reading comprehension. Next, the results of gender DDF discovered that the options in reading comprehension and vocabulary sections were more popular among female candidates, whereas the choices in the structure section were more favorite among males. Lastly, the results of discipline DDF identified that the options in reading

comprehension and vocabulary sections were popular among the HSS and N-HSS groups equally, whereas the options in the structure section were more popular among the HSS group. In conclusion, the findings of the present study would be worthwhile for the field of test validity and fairness, which can be noticeably improved through hybridizing DIF and DDF analytical methods.

7. Suggestions for Further Research

Based on the limitations of this study, a set of recommendations for further research can be provided. For example, some detailed studies could investigate the impact of the test format. The investigators could profitably compare the findings from multiple-choice tests to those from other test formats, such as open-ended questions or performance-based assessments. This can help determine if there are specific biases associated with certain test designs and provide insights into how to diminish them. Comparing the performance of different statistical methods for detecting DIF and DDF, such as the logistic regression approach is another suggestion for future studies. Moreover, the scholars are respectfully recommended to incorporate qualitative methods, such as interviews or focus groups, to gather in-depth insights into participants' experiences and perceptions related to gender and major differences in test performance. This can provide a richer understanding of the underlying factors contributing to DIF. Finally, the present study recommends that further research could examine DIF and DDF over time to determine if they persist or change over time. Replicating the study in different times, settings, and populations would strengthen the external validity of the results. This would allow for a more comprehensive understanding of the generalizability of the findings across different contexts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding

The authors received no financial support for the research, authorship, and publication of this article.

References

Adibatmaz, F. B. K., & Yildiz, H. (2020). The effects of distractors to differential item functioning in Peabody picture vocabulary test. *Journal of Theoretical Educational Science*, *13*(3), 530-547. https://doi.org/10.30831/akukeg.622180

Amirian, S. M. R. (2020). Investigating fairness of reading comprehension section of INUEE: Learner's attitudes towards DIF sources. *International Journal of Language Testing*, *10*(2), 88-100. https://doi.org/10.21437/ijlt.v10i2.1206

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *International Journal of Language Testing*, *4*(2), 187-203.

Baghaei, P. (2021). *Mokken scale analysis in language assessment.* Münster, Germany: Waxmann Verlag.

Baghaei, P., & Dourakhshan, A. (2016). Properties of single-response and double-response multiple-choice grammar items. *International Journal of Language Testing*, *6*(1), 33-49. URL: https://www.ijlt.ir/article_114425.html

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, *19*(1), 155–168. https://doi.org/10.31234/osf.io/9m5jw

Bakytbekovich, O. N., Mohammed, A., Alghurabi, A. M. K., Alallo, H. M. I., Ali, Y. M., Hassan, A. Y., Demeuova, L., Viktorovna, S. I., Nazym, B., & Afif, A. K. N. S. (2023). Distractor analysis in multiple-choice items using the Rasch model. *International Journal of Language Testing*, *13*(1), 69-78. https://doi.org/10.22034/ijlt.2023.387942.1236

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the subject area: A study of the Iranian national university entrance exam. *The Journal of Teaching Language Skills*, *2*(3), 1-26. https://doi.org/10.22099/jtls.2012.413

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3-20. https://doi.org/10.1177/0265532214531254

Berman, R. A., & Slobin, D. I. (Eds.). (2013). *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 157-64. https://doi.org/10.1111/j.1745-3984.1990.tb00740.x

Bortolotti, S. L. V., Tezza, R., de Andrade, D. F., Bornia, A. C., & de Sousa Júnior, A. F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity*, *47*, 2341-2360. https://doi.org/10.1007/s11135-012-9684-5

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422-433. https://doi.org/10.3758/BF03214546

Deng, J. (2020). *The Relationship between Differential Distractor Functioning (DDF) and Differential Item Functioning (DIF): If DDF Occurs, Must DIF Occur?* (Doctoral dissertation, University of Kansas).

Donlon, A. (1973). Content factors in sex differences on test questions. NJ: Educational Testing Service, *Research Memorandum*, *73*(1), 1-14.

Estaji, M., & Zhaleh, K. (2020). Does field of study matter in academic performance: Differential item functioning analysis of a high-stakes test using one-parameter and two-parameter item response theory models. *Iranian Journal of English for Academic Purposes*, *9*(3), 14-31. https://doi.org/10.22051/jap.2020.25663.2405

Fayers, P. M., & Machin, D. (2007). Item response theory and differential item functioning. *Quality of Life the Assessment, Analysis and Interpretation of Patient-Reported Outcomes, 2nd ed Chichester: Wiley*, 161-88.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika*, *60*(4), 459-487. https://doi.org/10.1007/BF02294324

Giguère, G., Bourassa, C., & Brouillette-Alarie, S. (2023). Effect of the differential item functioning (DIF) of LS/CMI items with convicted men and women. *Journal of Experimental Criminology*, *13*(2), 1-25. https://doi.org/10.1007/s11292-023-09559-9

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*(2), 147-160. https://doi.org/10.1111/j.1745-3984.1989.tb00325.x

Gregson, N., & Crang, M. (2010). Materiality and waste: Inorganic vitality in a networked world. *Environment and Planning A: Economy and Space*, *42*(5), 1026-1032. https://doi.org/10.1068/a43176

Halpern, D. F. (2013). *Sex differences in cognitive abilities*. Psychology Press. https://doi.org/10.4324/9780203816530

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media. https://doi.org/10.1007/978-94-017-1988-9

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*(5220), 41-45. https://doi.org/10.1126/science.7604277

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, N.J.: Erlbaum.

Huang, T. W., Wu, P. C., & Mok, M. M. C. (2022). Examination of gender-related differential item functioning through Poly-BW indices. *Frontiers in Psychology*, *13*, 821459. https://doi.org/10.3389/fpsyg.2022.821459

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53-69. https://psycnet.apa.org/doi/10.1037/0033-2909.104.1.53

Jamalzadeh, M., Lotfi, A. R., & Rostami, M. (2021). Assessing the validity of an IAU general English achievement test through hybridizing differential item functioning and differential distractor functioning. *Language Testing in Asia*, *11*(1), 1-17. https://doi.org/10.1186/s40468-021-00124-7

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59-76.

Koon, S., & Kamata, A. (2013). An applied examination of methods for detecting differential distractor functioning. *International Journal of Quantitative Research in Education*, *1*(4), 364-382. https://doi.org/10.1504/IJQRE.2013.058306

Lane, S., Wang, N., & Magon, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Researcher*, *15*(4), 2 1-27. http://dx.doi.org/10.1111/j.1745-3992.1996.tb00575.x

Linacre, J. M. (2009). *WINSTEPS Rasch Measurement* (Version 3.73) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2023b). Winsteps® Rasch measurement computer program User's Guide. Version 5.6.0. Portland, Oregon: Winsteps.com.

Maccoby, E., & Jacklin, C. N. (1974). Myth, reality and shades of gray: What we know and don't know about sex differences. *Psychology Today*, *8*(7), 109–112. http://dx.doi.org/10.1037/e400662009-008

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex-related performance differences on constructed-response and multiple-choice sections of advanced placement examinations. *College Board Report*, *92*(7), 1-37. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01516.x

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Metu, I. C. (2020). Using Rasch model to identify differential item functioning of teachers' job satisfaction scale with respect to gender. *The African Journal of Behavioural and Scale Development Research*, *2*(2), 59-66.

Middleton, K., & Laitusis, C. C. (2007). Examining test items for differential distractor functioning among students with learning disabilities. *ETS Research Report Series*, *2007*(2), i-34. https://doi.org/10.1002/j.2333-8504.2007.tb02085.x

Oga-Baldwin, W. Q., & Fryer, L. K. (2020). Girls show better quality motivation to learn languages than boys: Latent profiles and their gender differences. *Heliyon*, *6*(5), e04054. https://doi.org/10.1016/j.heliyon.2020.e04054

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*(3), 151-165. https://doi.org/10.1177/0146621609359284

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Studies in Applied Linguistics and TESOL*, *6*(2), 1-3. https://doi.org/10.7916/salt.v6i2.1548

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The University of Chicago Press, 1980). https://doi.org/10.7208/chicago/9780226702764.001.0001

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

Ravand, H., Rohani, G., & Firoozi, T. (2019). Investigating gender and major DIF in the Iranian national university entrance exam using multiple-indicators multiple-causes structural equation modelling. *Issues in Language Teaching*, *8*(1), 33-61. https://doi.org/10.22054/ilt.2020.49509.460

Salehi, M., & Tayebi, A. (2012). Differential item functioning: Implications for test validation. *Journal of Language Teaching & Research*, *3*(1), 84-92. https://doi.org/10.4304/jltr.3.1.84-92

Sandersfeld, T. J. (2020). *Differential item and distractor functioning between computer-based and paper-and-pencil testing within demographic groups on a Statewide Mathematics Assessment* (Doctoral dissertation, The University of Iowa). https://doi.org/10.17077/etd.9g6z8wmv

Shahmirzadi, N. (2023). Validation of a language center placement test: Differential item functioning. *International Journal of Language Testing*, *13*(1), 1-17. https://doi.org/10.22034/ijlt.2022.336779.1151

Shaywitz, B. A., Shaywltz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., Bronen, R. A., Fletcher, J. M., Shankweiler, D. P., Katz, L., & Gore, J. C. (1995). Sex differences in the functional organization of the brain for language. *Nature*, *373*(6515), 607-609. https://doi.org/10.1038/373607a0

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research*, *57*(2), 184-208. https://doi.org/10.1016/S0148-2963(01)00302-2

Willingham, W. W. & Cole, N. S. (1997). Fairness issues in test design and use. In Willingham, W.W. & Cole, N. S. (Eds.), *Gender and fair assessment*, (pp. 227 - 346). Hillsdale, NJ: Lawrence Erlbaum. http://www.questia.com/PM.qst?a=o&d=99105978