International Journal of Foreign Language Teaching and Research
ISSN: 2322-3898-http://jfl.iaun.ac.ir/journal/about
© 2022- Published by Islamic Azad University, Najafabad Branch

IJFLTR
International Journal of Foreign Language Teaching and Research

Please cite this paper as follows:
Jamalzadeh, M., Lotfi, A. R., & Rostami, M. (2022). Equity on General English Achievement Tests through Gender-based DIF Analysis across Different Majors. *International Journal of Foreign Language Teaching and Research, 10* (43), 47-65. http://doi.org/10.30495/JFL.2022.697333

Research Paper

Equity on General English Achievement Tests through Gender-based DIF Analysis across Different Majors

Mehri Jamalzadeh[1], Ahmad Reza Lotfi[2]*, Masoud Rostami[3]
[1]Ph.D. Candidate, Department of English language, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran
m82jamalzadeh@yahoo.com
[2]Assistant Professor, Department of English Language, Isfahan (Khorasgan), Branch, Islamic Azad University, Isfahan, Iran
lotfi.ahmdrzlotfi@gmail.com
[3]Assistant Professor, Department of Languages and Literature, Yazd University, Yazd, Iran
mrostami@yazd.ac.ir

Received: October 18, 2022 **Accepted: November 30, 2022**

Abstract
This study is an investigation of gender equity in the context of the General English Achievement Test developed and used at Islamic Azad University (Isfahan Branch, IRAN), henceforth IAUGEAT, with test takers majoring in different fields of study. A sample of 835 students sitting for IAUGEAT was chosen purposively. The test scores were analyzed by the one-parameter IRT model. A focus group interview (10 test developers and language teachers) was also used to inquire into their perceptions about the impact of test takers' gender and major on test equity. The findings of the DIF analysis indicated a reciprocal action between item type and gender DIF as some items exhibited DIF across different subgroups. In three subgroups, they favored female students. In one subgroup, they favored males. In the other two subgroups, they favored males and females alike. The results were further confirmed by the qualitative data obtained from the focus group interview. In general, our findings strongly suggest that checking gender equity via a Rasch-model DIF analysis is both eminent and convergent with a qualitative evaluation of test-takers' performance by test-developers and instructors.

Keywords: *Differential Item Functioning (DIF), Equity, Gender, General English achievement test, IAUGEAT, IRT, Test validation.*

برابری در آزمون های پیشرفت عمومی انگلیسی از طریق تجزیه و تحلیل DIF مبتنی بر جنسیت در رشته های مختلف
این پژوهش، بررسی برابری جنسیتی در چارچوب آزمون پیشرفت زبان انگلیسی عمومی است که در دانشگاه آزاد اسلامی (واحد اصفهان، ایران)، از این پس IAUGEAT، با آزمون‌دهندگان رشته‌های تحصیلی مختلف تهیه و مورد استفاده قرار می‌گیرد. نمونه ای از ۸۳۵ دانش آموز نشسته برای IAUGEAT به صورت هدفمند انتخاب شد. نمرات آزمون توسط مدل IRT یک پارامتری مورد تجزیه و تحلیل قرار گرفت. یک مصاحبه گروهی متمرکز (۱۰ برنامه‌نویس آزمون و معلمان زبان) نیز برای بررسی ادراکات آنها در مورد تأثیر جنسیت و رشته شرکت‌کنندگان در برابری آزمون استفاده شد. یافته‌های تجزیه و تحلیل DIF نشان‌دهنده یک عمل متقابل بین نوع آیتم و جنسیت DIF است زیرا برخی از آیتم‌ها DIF را در زیر گروه‌های مختلف نشان می‌دهند. در سه زیر گروه، دانشجویان دختر را ترجیح دادند. در یک زیر گروه، آنها مردان را ترجیح می دادند. در دو زیرگروه دیگر، آنها مردان و زنان را به طور یکسان ترجیح می دادند. نتایج بیشتر با داده های کیفی به دست آمده از مصاحبه گروهی متمرکز تأیید شد. به طور کلی، یافته‌های ما قویاً نشان می‌دهد که بررسی برابری جنسیتی از طریق تحلیل مدل Rasch با ارزیابی کیفی عملکرد شرکت کنندگان در آزمون توسط توسعه دهندگان آزمون و مدرسان، برجسته و همگرا است.

واژگان کلیدی: عملکرد آیتم دیفرانسیل(DIF) ، برابری، جنسیت، آزمون پیشرفت عمومی انگلیسی، IAUGEAT، IRT، اعتبار سنجی آزمون

Introduction

Validity is generally considered a yardstick for measuring test effectiveness. Test developers, consequently, focus on the validation procedure to guarantee the consequential aspects underlying the appropriacy of the right test or assessment for producing consistent results over time (for a recent validity argument in language testing, study Chapelle & Voss, 2021 and Winke & Brunfaut, 2021). It is greatly emphasized that the aftereffects evoked by a particular assessment or measure must not have any social and societal consequences (Messick, 1998). Therefore, given the pivotal role of tests used in education and their consequences for both the test takers and society, all experts in testing and assessment place a high premium on test developers’ efforts to improve the conditions required for promoting test validity and validation. Essentially, test score interpretation must be valid because drawing conclusions based on test scores is a critical issue, so evidence of test validity is highly recommended by language testing and assessment practitioners (Bejar, 1990; Chapelle, 1999; Embretson, 1994; Kane & Mislevy, 2017).

Every year about two thousand students in IAU (Isfahan Branch) take part in general English classes and are given the test. Yet, little attempt has been made to standardize a test bank. Other departments expect the best possible service, and in order to fulfill this expectation and improvement of IAUGEAT, the current study was done. Although there have been scientific and operational opportunities for standardization of these tests, they have not gone through the standardization procedure yet. So most probably, there is a high capacity for promoting the level of general English assessment and the related instruction and education. Thus in an attempt toward further standardization of IAUGEAT, the present study investigated whether IAUGEAT demonstrates significant DIF in favour of a certain gender group with different academic backgrounds.

Literature Review

Test validation

Test validation plays a pivotal role in the analysis of test fairness, and attempts to improve test validation or prevent test unfairness has gained considerable momentum in recent years. Notably, the fact that a measure must measure what it purports to measure has long been recognized as an essential element in testing of language (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2011; Im & McNamara, 2017; Kane, 1992, 2006, 2013; Lado, 1961; Messick,1989; Nakatsuhara, Taylor & Jaiyote 2018).

Score meaning and the value suggestions of scores serve as one of the major indicators that signal test validation. A unified perspective of test validity, therefore, requires a thorough understanding of scientific and ethical influences which govern test interpretation and use, whereby operational considerations such as content, criteria, and consequences need to be carefully considered. This can help test developers enhance the validity of the targeted tests since factors like appropriateness and usefulness of interpretations test score can guarantee trustworthiness of construct validity (Kane, 2006; Messick, 1995). In the classical model of test validity, construct validity is considered as significant validity evidence along with content validity and criterion validity. It is related to the extent a test evaluates what it intends to be evaluating. Test validation and test fairness theories place a high premium on the evidential bases supporting the interpretations of test scores. Accordingly, as Bachman and Palmer (2010) rightly argued, without a robust validation, it would be difficult to justify any decision made on the basis of test scores' interpretation. As a result, it should come as no surprise that the creation of validity theories in language testing has become one of the essential aspects of educational measurement (Kane, 2006; Messick, 1989).

Among validity theories in language testing, Mesick’s unitary notion of validity proposes two different but complementary groups of differences required to explain the characteristics of a given language test. The first group of distinctions deal with the proofs located within the construct being assessed, and the relatedness between the usefulness of the test and the construct in question. By contrast, the second group of distinctions are in relation to the test consequences accommodating the sources affecting the construct being measured, and the impact of the test use on both the test takers and the society. Not surprisingly, the validity theory established in terms of test-takers’ performance failed to address the practical guidelines governing test validation. Consequently, Kane (2006, 2012) proposed an interpretive-argumentative approach to test validation whereby “the logic, evidence, and rhetoric of arguments for the validity of an assessment” are of primary importance (Cumming, 2013, p. 3).

Developing fair and unbiased general English tests for university students from different majors is an issue of great importance. Thus domestic students studying various fields of study in non-English speaking countries might encounter problems understanding English texts and lectures because of an absence of English proficiency (Ramsay, Barker, & Jones, 1999). It has been argued that the main culprit might be related to university students’ failure to cope with the socio-cultural and psychological dimensions of their English proficiency. As a result, they lack the necessary English self-confidence in the procedure of sociocultural and psychological adjustment to academic situations involving the English language (Trice, 2007; Yang, Noels, & Saumure, 2006).

DIF Studies

These problems may be confounded when the targeted achievement tests do not evaluate what they claim to evaluate due to both internal and external factors. According to Wright (2007), both standardized or non-standardized achievement tests are used to determine what a student has learned, such as vocabulary, grammar, and reading by specifying how much of the teaching content has been mastered by the targeted learners. What makes such tests really significant is the fact that they are designed to measure learners’ current levels of knowledge for helping them advance at a suitable pace. Consequently, validation, equity, and fairness must carefully be considered in general English achievement tests. The European Federation of Psychological Association (E.F.P.A) has recently emphasized the necessity of collecting evidence attesting to the construct validity of high stake tests (Hope, Adamson, McManus, Chris, & Elder, 2018) and has maintained that differential item functioning (DIF) as an effective method of evaluating test quality for measuring the quality of the test, should be used because DIF analysis has a great bearing on test equity and fairness (Hernández, Tomás, Ferreres, & Lloret, 2015).

Accordingly, carefully designed quantitative approaches are often utilized to specify whether the test items or test scores have equivalent meaning for varying groups of test-takers. In other words, test fairness and equity are important considerations, so test developers must focus on the examinees’ background characteristics like gender in order to guarantee the validity of test scores. To this end, differential item functioning (DIF) techniques are used to find biased items that have an adverse effect on the test validity and may lead to the unfair evaluation of test-takers’ performance with different personal characteristics but the same language ability (Ozemir & Alshamrani, 2020).

Mentioning some studies investigating DIF in high stake tests, Barati and Ahmadi (2012) did research on DIF on the Special English Test of the Iranian National University Entrance Exam. Grammar, language function, and cloze sections favored females, whereas vocabulary and word order sections favored males. Both men and women were favored on the reading comprehension section equally. Alavi and Bordbar (2018) investigated gender DIF in language proficiency test in Iran, the National University Entrance Exam for Foreign Languages. The results showed that 40 out of 95 items of the test exhibited DIF, suggesting that the test marks are not free of construct-irrelevant variance. Darabi, Bazvand, and Ahmadi (2020) examined test fairness, focusing on the subject-matter section of the Ph.D. The Entrance Exam of ELT held in 2014 in Iran. The results underscored that the test is biased since the tasks were not fully discussed in the Ph.D. course objectives, the test was best reliable for high-ability test-takers and 4 items were flagged for nonnegligible DIF.

In another related study, focusing on the validity of a General English Achievement Test, Jamalzadeh, Lotfi and Rostami (2021) examined both DIF and differential distractor function (DDF) items. The findings revealed five moderate-level DIF and ten DDF items indicating an adverse effect on test fairness. Bordbar (2021) explored the validity and DIF analytics of Iran's University Entrance Exam. The test results revealed that the test marks were not without construct-irrelevant variance and the test's fairness was not clarified. Mehrazmay, Ghonsooly, and De La Torre (2021) examined gender DIF in the 15-item reading comprehension section of the university entrance exam of ELT held in 2017 in Iran. Three items displayed large DIF. The findings show that women have lower chance of correct answer across all latent profiles.

Several statistical procedures for finding DIF have been proposed. For instance, the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), SIBTEST (Shealy & Sout, 1993), Item Response Theory (IRT) methods (Camilli & Shepard, 1994), logistic regression (Swaminathan & Rogers, 1990), and multilevel DIF analysis (Kamata, 2001) are commonly used by the researchers' interested item analysis in language testing. Khodi and Karami (2021) investigated the comparability of findings from three extensively used DIF finding techniques: the Rasch model, Logistic Regression, and Mantel-Haenszel through the data from 35 item grammar section of the University of Tehran English Proficiency Test. This study, however, employed the MH, which can be employed for comparing two cultural groups when the observed item scores are dichotomous and the sum score represents the targeted latent variable.

Purpose of the Study

General English tests are commonly classified as high-stakes tests because the marks are often used as crucial indicators in determining students' future academic accomplishments. In the Iranian academic scene, a large number of students take part in general English classes and are tested at the end of the semester. Evidently, no serious attempt has been made to standardize the general English test banks. Although there have been many scientific and operational opportunities for standardization of these tests, they have not gone through the standardization procedure considering viral factors like the role of test takers' gender and major type on the quality of test equity and test fairness. The significance of investigating test equity by identifying any sources of bias in different academic contexts is vitally important in language testing where interpretation of test scores may have a great bearing on students' educational and professional opportunities. Unfortunately, in the majority of studies addressing test equity, the concerned practitioners have neglected the cumulative impact of both gender and major type on the validity of general English achievement tests.

Thus, in an attempt toward further standardization of IAUGEAT, the present study investigated whether IAUGEAT shows substantial DIF in favour of gender groups with specific majors. To this purpose, the following research questions were raised:

RQ$_1$: *To what extent does the examinees' field of study influence any possible gender-based DIF?*

RQ$_2$: *What are the attitudes and subjective analyses of any possible gender bias in the test according to language instructors and educators at IAU?*

Method

With adopting a convergent mixed parallel design method integrating quantitative and qualitative gathering, grouping, interpreting, and analyzing the related data, an attempt was made to examine the items exhibiting potential gender DIF in IAUGEAT. The main purpose of this design is to set up the requirements needed to best understand or develop a more complete picture of the research problems by obtaining different but complementary data. In fact, the method helps collect and analyze two independent strings of quantitative and qualitative data simultaneously in a single phase. By prioritizing the methods equally, maintaining the data analysis independently, and mixing the results during the overall interpretation, the method paves the way for discovering convergence, divergence, contradictions, or relationships of two sources of data.

In the quantitative stage, Mantel- Haenszel DIF detection was applied to the data related to different subgroups under scrutiny. In the qualitative stage, however, a focus group interview was used where a limited number of EFL teachers were randomly selected from the sample teachers teaching general English during the semester were randomly selected and asked about their opinion or perceptions about the validity and fairness of the test. By creating an interactive environment where the participants were able to freely discuss the possible causes of DIF and suggest solutions for improving the items on the test.

Participants

Test Takers

The data for this study was collected from 835 male and female undergraduate students in different majors at Isfahan Azad University (IAU), Iran. They had all sat for the IAUGEAT administered in the fall of the 2018-2019 academic year. Table 1 presents the participants’ major. Notably, 63% of the test takers were females and 37% were males.

Table 1

Test Population in Terms of Gender

Subgroup	Total	Fe male	Male
Para medicine	80	55	25
Educational Sciences	227	197	30
Humanities	148	78	70
Agriculture	143	92	51
Physical Education	45	24	21
Engineering	109	31	78
Architecture	83	47	36
All Subgroups	835	524	311

Focus Group Participants

The researcher asked the language instructors and educators at the English department at IAU, Isfahan branch, to take part in the focus group and share their ideas on the possible cause of DIF and suggest ways to improve the items. Only ten of them participated in the focus group. Details of the focus group are summarized in Table 2.

Table 2

Focus Group Details

No.	Female	Male	Age	Education	Teaching Experience
10	5	5	35-50	Ph.D.	10-25 years

Data Collection

In general, the students taking IAUGEAT answer the questions manually. After grading the answer sheets, the teachers deliver them to the examination office. We obtained official permission from the university authorities in order to use the answer sheets as their source of data on the condition of confidentiality exclusively for this research project.

Instruments

IAUGEAT

Generally, students are required to take a general English course and pass the related exam as one of the mandatory modules in the undergraduate program. The assessment tool used for measuring students' knowledge of English is a general English achievement test. Therefore, the IAUGEAT used for assessing students from different majors in the autumn semester of 2018-2019 academic year was the instrument chosen for the DIF analysis. The test was organized into four sections: Vocabulary (25 questions), Grammar (15 questions), Cloze Test (10 questions), and Reading Comprehension (10 questions).

We employed jMetrik software in order to appraise the reliability from a single test administration. It computed Huynh's raw agreement and kappa statistics. Table 3 summarizes the result of the analysis.

Table 3

Huynh's Raw Agreement Index Statistics

Vocabulary	Grammar	Cloze Test	Reading	Total
0.90	0.81	0.82	0.90	0.92

Clearly, the grammar section of the test has the lowest value of reliability. Vocabulary and reading comprehension parts, however, have the highest reliability values. Overall, the test total reliability is equal to 0.92, showing an acceptable level of reliability.

Focus Group Interview

As already mentioned, ten of the language instructors handling general English courses at IAU, Isfahan branch, volunteered to participate in the focus group interview. The main objective was to find out their opinions on the possible causes of DIF and suggest ways to improve the test items. The information concerning the participating language teachers for the focus group interview is summarized in Table 4.

Data Processing

Detecting DIF

DIF exists when one specific group of test takers has a distinct expected item score than similar test takers from a different group. The condition suggests that certain items are evaluating something beyond the purported construct (Angoff, 1993; Pae, 2012; Meyer, 2014). Osterlind (1983) introduced five procedures for detecting possible bias in test items: By analysis of variance (ANOVA), transformed item difficulties (TID), chi-square (χ2), item characteristic curve (ICC), and distractor response analysis. Chi-square (χ2) was used as the main strategy for detecting possible test item bias in this study. To do so, jMetrik software was applied for psychometric analyses. First, 835 test takers' scores were loaded into an excel file. Then, they were converted into a notepad file required for jMetrik. Additionally, for testing statistical significance, common-odds ratio, ETS delta statistic, and the standardized mean difference

(SMD) for describing practical significance, Cochran-Mantel-Haenszel (CMH) statistic was utilized. Notably, the analysis was separately done for each of the subgroups.

Assumptions Underlying Rasch model

There are two statistical suppositions; unidimensionality and local independence. Unidimensionality signifies that the test consists of items that make use only of one dimension. In other words, in this model, there is a single Ɵ for each testee, and other factors that might be influencing the item response are treated as a random error or nuisance dimensions. These factors are considered to be item-specific in that every item is unique and independent from other test items (DeMars, 2010).

A principle specific to Rasch model is that the comparison between the characteristics of any two testees should be equal no matter what subset of items is used for the comparison, and the comparison between the properties of any two items should be equivalent without regard to which subset of individuals is used for the comparison. This principle is known as specific objectivity. (Andrich & Marais, 2019; Bond & Fox, 2013; Rasch, 1977). In order to meet the assumption of uni-dimensionality in the Rasch model, the test was divided into its different parts and the analysis was repeated for each of these parts independently.

Alternatively, the assumption of local independence is, in reality, a provision in jMetrik which acts as an option for checking the assumption of local independence with Yen's *Q*3 statistic, defined as the correlation of residuals for a pair of items (Yen, 1984, cited in Meyer, 2014). By performing a correlation analysis, it was found out that no extreme values were present and the supposition of local independence was also supported.

CMH Chi-Square Statistic

Cochran–Mantel–Haenszel test (CMH) is a test employed in the investigation of stratified or matched categorical data. It helps the researcher to test the connection between a binary predictor and a binary outcome such as case or control status while taking into conideration the stratification. This procedure aims to test the formulated null hypothesis based on item scores providing every examinee is an independent member of the targeted group. By stratifying examinees in terms of matched scores and evaluating the difference between the observed and expected item scores pooled over all strata, it is possible to test the target hypothesis. In case the difference existing between matched scores is not due to chance factors, the null hypothesis will be rejected.

ETS DIF Classification Levels

According to Meyer (2014), ETS DIF classification levels could be described by ETS rules regarding the common odds ratio related to the magnitude of DIF. There are three types of rules labeled as A items, B items, and C items. "A" items have a CMH p-value larger than 0.05, that means the common odds ratio is strictly between 0.65 and 1.53. "B" items, on the other hand, have a common odds ratio lower than 0.53 and the upper bound of the 95% confidence interval for the common odds ratio is less than 0.65. That is, the common odds ratio is larger than 1.89 and the lower bound of the 95% confidence interval is larger than 1.53.

Results

In the quantitative phase, the data related to the comparative research whose main objective was to grant a comprehensive representation of the evidential bases regarding the targeted research questions, will be displayed and then will continue with a qualitative phase whereby an interview was used to further explore the importance of equity and test fairness in the interpretation of test scores.

Results of the quantitative stage

DIF analysis was used for each subgroup of testees from different majors taking the test. In particular, each of the Agriculture and Engineering subgroups had one item classified as "C" showing a large amount of DIF. In the Agriculture subgroup, item 1 was qualified as C+, but in the Engineering subgroup item, 30 was qualified as C+. Notably, in other subgroups, certain items showed B magnitude of DIF which will be explained below.

Educational Sciences DIF analysis evaluated eight items (13.33%) showing a moderate magnitude of DIF. Four items (i.e., items 10, 17, 19, and 34) were classified as B- and was in the favor of the reference group (i.e., female students) (6.66%), whereas items 21, 42, 43, and 51 were classified as B+ and favored the focal group or male students (6.66%). The Characteristics of these items are shown in Table 4.

Table 4

Items exhibiting DIF in Educational Sciences

Item No.	Subtest	Chi-Square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
10	V	4.46	0.03	B-	0.64	-0.31	0.46
17	V	6.43	0.01	B-	0.60	-0.09	0.24
19	V	4.22	0.04	B-	0.66	-0.42	0.63
21	V	9.38	0.00	B+	0.82	-1.47	0.47
34	G	7.85	0.01	B-	0.44	0.71	0.30
42	C	4.40	0.04	B+	0.52	0.33	0.44
43	C	4.93	0.03	B+	0.67	-0.45	0.50
51	R	3.86	0.05	B+	0.74	-0.90	0.41

As indicated in Table 5 the highest percentage of DIF for the subgroup in Educational sciences belongs to the cloze test and favors females. Similarly, the Reading section of the test also favors females. The vocabulary and grammar parts of the test, however, favor males. The cloze and reading comprehension parts were basically related to comprehension check and female students in Educational sciences outperformed males in answering contextualized language items which need a more holistic view of the text. However, males were good at decontextualized language items.

Table 5

Percentage of DIF in Different Test Parts in Educational Sciences Subgroup

Test Part	Vocabulary	Grammar	Cloze Test	Reading
DIF Favoring Males	12%	6.66%	0%	0%
DIF Favoring Females	4%	0%	20%	10%
Total DIF	16%	6.66%	20%	10%

In the Para-medicine subgroup DIF analysis, six items (10%) exhibited moderate DIF magnitude. Five items were qualified as B- favoring male students (8.33%) and one item qualified as B+ favoring female students (1.66%). Items 2, 18, 21, 44, and 51 were classified as B- favoring the reference group (i.e., males), and item 29 was identified as B+ favoring the focal group or females. The characteristics of these items are depicted in Table 6.

Table 6

Items exhibiting DIF in Para-medicine Subgroup

Item No.	Subtest	Chi-Square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
2	V	4.77	0.03	B-	0.58	-0.02	0.47
18	V	3.19	0.01	B-	0.95	-1.35	0.30
21	V	5.27	0.05	B-	0.99	-2.83	0.005
29	G	4.59	0.03	B+	0.81	0.23	0.30
44	C	4.57	0.03	B-	0.75	0.63	0.45
51	R	3.85	0.02	B-	1.00	-4.05	0.30

It is clearly observed that in the para-medicine subgroup, the highest percentage of DIF belongs to the vocabulary section and is in favor of males. The results related to both cloze test and reading sections favor males, while the grammar part favours students. It seems that the items favoring males or females change for different subgroups. Likewise, Table 7 provides a summary of the DIF analysis for the para-medicine subgroup.

Table 7

Percentage of DIF in Different Test Parts in Para-medicine Subgroup

Test Part	Vocabulary	Grammar	Cloze Test	Reading
DIF Favoring Males	12%	0%	10%	10%
DIF Favoring Females	0%	6.66%	0%	0%
Total DIF	12%	6.66%	10%	10%

The analysis of the test items in the Physical Education subgroup identified four items (6.66%) as exhibiting a moderate magnitude of DIF. Two items were in favor of male students (3.33%) and two items were in favor of females (3.33%). Items 45 and 55 were qualified as B-, favoring the reference group (i.e., the male students), while items 33 and 57 qualified as B+ favoring the focal group or females. The characteristics of these items are displayed in Table 9.

Table 8

Items exhibiting DIF in Physical Education Subgroup

Item No.	Subtest	Chi-square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
33	G	7.05	0.01	B+	0.89	-0.16	0.30
45	C	3.85	0.05	B-	0.47	1.29	0.31
55	R	5.12	0.02	B-	0.76	1.07	0.59
57	R	6.10	0.01	B+	0.87	0.10	0.30

In the Physical Education subgroup, the highest percentage of items exhibiting DIF belonged to the reading test section favoring both males and females equally. The vocabulary part showed no DIF, whereas the grammar part favored females, and the cloze test favored males. Percentages of items exhibiting DIF in the Physical Educational subgroup for each test part for males and females are depicted in Table 9.

Table 9

Percentage of DIF in Different Test Parts in Physical Educational Subgroup

Test Part	Vocabulary	Grammar	Cloze test	Reading
DIF Favoring Males	0%	0%	10%	10%
DIF Favoring Females	0%	6.66%	0%	10%
Total DIF	0%	6.66%	10%	20%

The analysis of the Architecture subgroup revealed that four items (6.66%) had DIF. One item favored male students (1.66), and three items favored female students (5%). Item 42 was qualified as B- and items 14, 51, and 53 as B+. The characteristics of these items are depicted in Table 10.

Table 10

Items Exhibiting DIF in Architecture Subgroup

Item No.	Subtest	Chi-square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
14	V	7.12	0.01	B+	0.58	0.37	0.49
42	C	3.91	0.05	B-	0.49	0.85	0.35
51	R	3.94	0.05	B+	0.79	-0.97	0.51
53	R	5.63	0.02	B+	0.79	-0.83	0.51

As can be seen, the highest percentage of items exhibiting DIF was related to the Reading part, and these items favored female students in general. While the Vocabulary part favored females, the cloze test favored males. The grammar part showed no DIF. All in all, the point that there is a lack of a fixed pattern for items favoring males and females was confirmed by going through the details of this subgroup too. Table 11 reports DIF percentages for the Architecture subgroup.

Table 11

Percentages of DIF in Different Test Parts in Architecture Subgroup

Test Part	Vocabulary	Grammar	Cloze Test	Reading
DIF Favoring Males	0%	0%	10%	0%
DIF Favoring Females	4%	0%	0%	20%
Total DIF	4%	0%	10%	20%

Similarly, the analysis of the Agriculture subgroup revealed three items (5 %) having DIF. Item1 was classified as C+ and items 4 and 37 were classified as B+. These items were in favour of the focal group or females. The characteristics of these items are depicted in Table 12.

Table 12

Items Exhibiting DIF in Agriculture Subgroup

Item No.	Subtest	Chi-square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
1	V	7.81	0.01	C+	0.83	-1.23	0.24
4	V	7.30	0.01	B+	0.58	0.33	0.34
37	G	4.39	0.04	B+	0.65	-0.08	0.51

Clearly, the highest percentage of DIF was in the vocabulary part favoring female students. The Grammar part also favored female students. Cloze test and reading parts showed no DIF. Once again, an absence of a fixed pattern for items favoring males and females was confirmed by going through the details of the subgroup.

Table 13
Percentage of DIF in Different Test Parts in Agriculture Subgroup

Test Part	Vocabulary	Grammar	Cloze Test	Reading
DIF Favoring Males	0%	0%	0%	0%
DIF Favoring Females	8%	6.66%	0%	0%
Total DIF	8%	6.66%	0%	0%

The analysis of the items for the Engineering subgroup indicated that two items (3.33%) exhibited DIF. One of the items was in favour of female students (1.66%) and the other one favored male students (1.66%). In fact, item 30 exhibited C magnitude of DIF and favored female students, and item 45, exhibiting B magnitude of DIF, favored the male students. The characteristics of these items are depicted in Table 14.

Table 14
Items Exhibiting DIF in Engineering Subgroup

Item No.	Subtest	Chi-square	P-value	Class	Item Difficulty (CTT)	Item Difficulty (Rasch)	Item Discrimination
30	G	8.99	0.00	C+	0.36	1.41	0.25
45	C	4.61	0.03	B-	0.42	1.11	0.05

It is clearly observed that in the Engineering subgroup, the highest percentage of DIF belonged to the cloze test, which favoured male students. The grammar part of the test favored female students only, while the vocabulary and reading parts exhibited no DIF. The result of the DIF analysis for the engineering subgroup is summarized in Table 15.

Table 15
Percentage of DIF in Different Test Parts in Engineering Subgroup

Test Part	Vocabulary	Grammar	Cloze Test	Reading
DIF Favoring Males	0%	0%	10%	0%
DIF Favoring Females	0%	6.66%	0%	0%
Total DIF	0%	6.66%	10%	0%

Clearly, DIF analysis of the test items belonging to the Humanities Subgroup demonstrated a negligible, A-level DIF. Table 16 presents the percentage of items favouring males and females.

Table 16
Percentage of Items Favoring Females and Males in Each Subgroup

Subgroup	Educational Sciences	Para-medicine	Physical Education	Architecture	Agriculture	Engineering	Humanities
Males	6.66%	8.33%	3.33%	1.66%	0%	1.66%	0%
Females	6.66%	1.66%	3.33%	5 %	5%	1.66%	0%

Total	13.32%	9.99	6.66%	6.66%	5%	3.32%	0%

Looking at the DIF analysis of the items answered by testees from different subgroups indicates that some of the items on the test under scrutiny favored males while others favored females. Overall, females were more favored than males in Grammar and Reading, while in the Vocabulary and Cloze subsections of the test, males were more favored than females. The related percentages are reported in Table 17.

Table 17

Number and Percentage of Items Favoring Females and Males in Each Test Part

Test Part	Females	Males	Females	Males	Total DIF
Vocabulary (25 items)	4	6	16%	24%	40%
Grammar (15 items)	4	1	26.6%	6.66%	33.26
Cloze Test (10 items)	2	4	20%	40%	60%
Reading (10 items)	4	2	40%	20%	60%

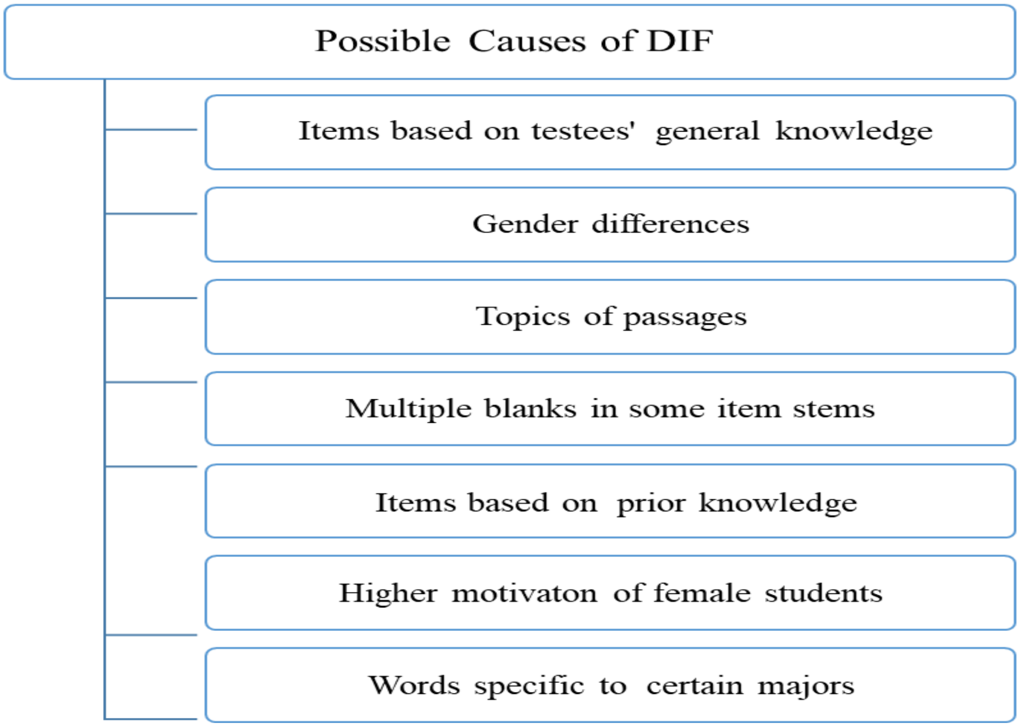
As can be seen, items comprising the targeted test defy test equity and fairness and behave differently for different genders from different major types.

Results of the qualitative stage

The results of the focus group interview are presented in Figure 1.

Figure 1

Results of the Focus Group interview


Possible Causes of DIF
Items based on testees' general knowledge
Gender differences
Topics of passages
Multiple blanks in some item stems
Items based on prior knowledge
Higher motivaton of female students
Words specific to certain majors

The interview directed under the guidance of the researcher brought to the surface a number of causes that adversely affected test fairness and equity. It was admitted that the construction of high-stakes general achievement tests required carefully designed plans considering all aspects of test bias.

The majority of the respondents who were interviewed felt that the targeted list of factors causing DIF might have a great bearing on the so-called rest equity and fairness. Comparing the results obtained from the two phases of the study, it was seen that the overall response to the questions posing the causes of DIF supported the experimental evidence overwhelmingly.

Discussion

This study aimed to apply DIF analysis to identify the possible threat to test validity by examining whether the items used in a given test function differentially among distinct sub-populations of the testees across different fields. We applied DIF analyses to find the biased items in a 60 item, multiple choice General English Achievement Test administered at IAU in Iran in order to examine the extent to which the examinees' field of study influence possible gender-based DIF. To achieve the predicted objectives, a two phase study plan comprising a quantitative and a qualitative phase was devised to investigate the validity of the test and its fairness. In the first phase, the main objective was to determine the extent to which the examinees' field of study influenced any possible gender-based DIF and to identify whether the items on the targeted test exhibited DIF in different subgroups or majors. The results of the quantitative phase, the gender-based DIF analysis of the test items, revealed that the examinees' field of study had a considerable effect on their overall test scores.

The findings are consistent with Ryan and Bachman (1992) who applied gender-based DIF analysis to examine the test performance of male and female test takers on TOEFL and FCE tests. The results revealed that four items favored males and two of them were biased toward females in the TOEFL test, whereas only one of the items favored males and one was in favor of the females in the FCE test.

The results were also in agreement with Barati and Ahmadi's (2012) findings based on gender-based DIF analysis showed that certain items in the grammar, language function, and cloze sub-sections of the Iranian National University Entrance Exam subtest favored females, while some of those in the vocabulary and word order sections favored males. It is encouraging to compare the results of the study with those found by Alavi and Bordbar (2018), who used gender-based DIF analysis to investigate test fairness in an Iranian high stake language proficiency test named the National University Entrance Exam for Foreign Languages (NUEEFL). The results revealed that 40 items out of the 95 items on the test contained irrelevant construct variance adversely affecting test equity.

The current findings also accord with the observations reported by Ravand, Firoozi, and Rohani in 2019. The results demonstrated that about half of the DIF items identified in the general English section of the university entrance exam for the English Master Program were contaminated with gender bias. Likewise, the findings of the study are consistent with those reported by Yoon (2020), and Geramipour (2020) who used DIF analysis reading comprehension and the findings of the studies indicated the presence of ten DIF items with a large size effect.

The thematic interpretation of the interview with the selected general English teachers based on the prespecified themes by the researcher in the second phase of the study demonstrated that there were possible sources of threat to validity and test fairness in the targeted general English test. During the focus group discussions, almost two-thirds of the participants (64%) pointed to and agreed with several causes of DIF. A common theme emerging out of the discussion was the case of topic familiarity for males in the cloze subsection in the test. Approximately half of those surveyed commented that the content of the text used in reading and cloze sub-parts of the test was responsible for the emergence of DIF. They stated that the questions in the cloze test could have been answered using general knowledge and they were biased towards males due to the text content and topic familiarity. They emphasized that the selection of content for reading comprehension and cloze subsections of the test should be chosen with utmost care and sensitivity.

Some items like item 1 in which there is a male proper name, or item 2 in which the item stem includes the word “brother” cause DIF because they point to a specific gender and such words or proper names could influence the perception and recall of things and events. Item 1 with the proper name” Reza” in the Agriculture subgroup analysis with a large magnitude of DIF functioned to the advantage of females. Similarly, item 2 with the word” brother” functioned to the advantage of males in the Para-medicine subgroup. Replacing the stem of such items using short dialogues with male and female speakers was suggested for removing DIF in such items. In response to the causes of DIF for items favoring females, 70 % of those interviewed said that the reason for the higher percentage of DIF favoring females was that test developers were also females and they enjoyed the same cognitive learning style and a common perspective on the world as the female testees. Another reason was that obtaining a higher score has been of higher importance for females and this likely led to putting more effort even in the exam session and has consequently caused more response validity for females.

Overall, the results obtained from the interpretation of the comments in the focus group interview revealed that the majority of those in the interview (90%) completely agreed with prespecified causes of DIF and suggested that there was a great need for finding more effective methods of constructing general English tests where the items are bias-free only measuring the construct under scrutiny. It is observed that the results of this study corroborate the findings of a great deal of the previous work in the field of testing, where the application of DIF analysis could practically improve the quality of test validation (Bank, 2009; Belzak & Bauer, 2020; Chen Liu & Zumbo, 2020; Mckeown & Oliveri, 2017; Paulsen *et al*. 2020; Zhu & Aryadoust, 2020, among many others).

Conclusion

The present study aimed at examining the effect of gender on test-takers seating for a general English achievement test at IAU. It was found that the items on the test exhibited DIF in different subgroups. It revealed that three subgroups favored female students more than male students (Architecture, Agriculture, and Humanities), however, one subgroup favored males more than females. (Para-medicine) and two subgroups favored males and females equally (Educational Sciences and Engineering).) Regarding the test parts, in Grammar and Reading, females were more favored than males, while in Vocabulary and Cloze Test males were more favored than females. The findings of this study may bring about certain implications regarding gender DIF in IAUGEAT. The results may be advantageous to test developers by providing information concerning the influence of gender on the performance of test-takers. By identifying items free of DIF and modifying or eliminating those violating the validity of test, test developers should have a bank of tests in which the items are purely bias-free. Considering the scarcity of the DIF research on general English achievement tests, the present research could be insightful to the practitioners in this field and used as a platform for further studies in this regard. The results of the present study could also be helpful to teachers and learners. The results of the study have important implications for developing bias-free general English achievement tests. Validity is a multifaceted phenomenon that should be the main focus of test construction. However, to clearly understand the true nature of the influence of gender and testees’ type of major on the interpretation of test scores, more research needs to be undertaken. More DIF studies focusing on the interaction effect of field of study with other factors can be very enlightening and will deepen understanding of the DIF and its possible causes. The evidence from this research suggests that more research is required to fathom out the true nature of test validity and test fairness. It was stated that of numerous methods available for detecting DIF, just one method was used in the present study for analysis and this could be regarded as one of the limitations of the study. The results of the current study were based on a limited item pool of 60 items; therefore, it needs to be repeated with larger samples of test items. Furthermore, the focus group was managed without the presence of the testees taking the targeted test. Their absence made the researcher's chance of reaping the benefits of the insider perspective, where those belonging to the study were not provided with the opportunity to express their own ideas. Taken together, the findings of this study pave the ground for implementing projects at the national level through the collaborative work of concerned researchers, shedding light on the factors causing item DIF and contaminating test fairness.

References

Alavi, S. M., & Bordbar, S. (2018). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. MOJES: *Malaysian Online Journal of Educational Sciences*, *5*(1), 10-24.

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. *Measuring in the Educational, Social and Health Sciences*, 41-53.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2(1)*, 1-34. https:/ /doi.or g/10.1 207/s1 54343 11laq0201_1

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Banks, K. (2009). Using DDF in a post hoc analysis to understand sources of DIF. *Educational Assessment*, *14*(2), 103-118.

Barati, H., & Ahmadi, A. R. (2012). Gender-based DIF across the subject area: A study of the Iranian National University Entrance Exam. *Journal of Teaching Language Skills, 29*(3), 1-26.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*(3), 237-245.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement*, *50*(1), 74-83.

Bond, T. G., & Fox, C. M. (2013). Applying the Rasch model: *Fundamental measurement in the human sciences*. Psychology Press.

Bordbar, S. (2021). Gender Differential Item Functioning (GDIF) Analysis in Iran's University Entrance Exam. *English Language in Focus (ELIF)*, *3*(1), 49-68.

Camilli, G., & Shepard, L. A. (1994). MMSS: Methods for identifying biased test items.

Camilli, G. (2018). IRT Scoring and Test Blueprint Fidelity. *Applied Psychological Measurement*, *42*(5), 393-400.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453-472.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language TM*. Routledge.

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity Argument in Language Testing: Case Studies of Validation Research*. Cambridge University Press

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

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics, 10*(4), 417-451.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1-8. https: //doi.o rg/10.108 0/154343 03.2011.22016

Darabi Bazvand, A., & Ahmadi, A. (2020). Interpreting the Validity of a High-Stakes Test in Light of the Argument-Based Framework: Implications for Test Improvement. *Research in Applied Linguistics*, *11*(1), 66-88.

DeMars, C. (2010). Item response theory. Oxford University Press.

Embretson, S. (1994). Applications of cognitive design systems to test development. In *Cognitive assessment* (pp. 107-135). Springer, Boston, MA. Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ:Author.

Educational Testing Service (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Geramipour, M. (2020). Item-focused trees approach in differential item functioning (DIF) analysis: a case study of an EFL reading comprehension test. *Journal of Modern Research in English Language Studies*, *7*(2), 123-147.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6(4), 269-279.*

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hernández, A., Tomás, I., Ferreres, A., & Lloret, S. (2015). THIRD EVALUATION OF TESTS PUBLISHED IN SPAIN. *Papeles del Psicólogo*, *36*(1), 1-8.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer H & Braun HI (Eds.), Test validity (pp. 129–145). Hillsdale, NJ, US.

Holland, P. W., & Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale NJ: Erlbaum.

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC Medical Education*, *18*(1), 1-7.

Im, G. H., & McNamara, T. (2017). Legitimate or illegitimate uses of test scores in contexts unrelated to test purposes. *English Teaching*, *72*(2), 71-99.

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/s40468-021-00n124-7.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kamata, A., & Vaughn, B. K. (2004). An Introduction to Differential Item Functioning Analysis. *Learning Disabilities: A Contemporary Journal*, *2*(2), 49-69.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112(3), 527.*

Kane, M. (2006). Content-related validity evidence in test development. *Handbook of Test Development*, *1*, 131-153.

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement:Interdisciplinary Research & Perspective*, 10(1-2), 66-70.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11-24). Routledge.https://doi.org/10.22059/jflr.2021.315079.783

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

Kunnan, A. J. (2004). Test fairness. *European language testing in a global context*, 18, 27-48.

Lado, R. (1961). *Language testing*. London: Longmans.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303), 690-700. https://doi.org/ 10.1080/01621459.1963.10500879

McKeown, S. B., & Oliveri, M. E. (2017). Exploratory analysis of differential item functioning and its possible sources in the National Survey of Student Engagement.

Mehrazmay, R., Ghonsooly, B., & De La Torre, J. (2021). Detecting Differential Item Functioning Using Cognitive Diagnosis Models: Applications of the Wald Test and Likelihood Ratio Test in a University Entrance Examination. *Applied Measurement in Education*, 1-23.https://doi.org/ 10.1g080/08957347.2021.1987906

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1), 35-44.

Meyer, J. P. (2014). Applied measurement with jMetrik. Routledge.

Nakatsuhara, F., Taylor, L., & Jaiyote, S. (2018). The role of the L1 in testing L2 English. Cambridge University Press.

Osterlind, S. J. (1983). *Test item bias* (No. 30). Sage.

Ozdemir, B., & Alshamrani, A. H. (2020). Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods. *International Journal of Learning, Teaching and Educational Research*, *19*(6), 27-45.

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, 29(4), 533-554. https: //doi.o rg/10.11 77/02655 32211434 027

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267-281.

Purpura, J. E. (2011). Quantitative research methods in assessment and testing. In *Handbook of research in second language teaching and learning* (pp. 749-769). Routledge.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research 1. *Language Learning*, *65*(S1), 37-75.

Ramsay, S., Barker, M., & Jones, E. (1999). Academic Adjustment and Learning Processes: a comparison of international and local students in first-year university. *Higher Education Research & Development*, *18*(1), 129-144.

Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements in symposium on scientific objectivity, Vedbaek, Mau 14-16, 1976. Danish Year-Book of Philosophy Kobenhavn, 14, 58-94.

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.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. *The Sage handbook of quantitative methodology for the social sciences, 107-116.*

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language testing*, *9*(1), 12-29.https://doi.org/ 10.1177/026553229200900103

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Stansfield, C. W., & Hewitt, W. E. (2005). Examining the predictive validity of a screening test for court interpreters. *Language Testing*, *22*(4), 438-462.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Trice, A. G. (2007). Faculty Perspectives regarding Graduate International Students' Isolation from Host National Students. *International Education Journal*, *8*(1), 108-117.

Willingham, W. W. (1999). A systemic view of test fairness. *Assessment in higher education: Issues of access, quality, student development, and public policy*, 213-242.

Winke, P., & Brunfaut, T. (Eds.). (2021). *The Routledge handbook of second language acquisition and language testing*. Routledge.

Wright, R. J. (2007). *Educational assessment: Tests and measurements in the age of accountability. Sage Publications.*

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

Yang, R. P. J., Noels, K. A., & Saumure, K. D. (2006). Multiple routes to cross-cultura adaptation for international students: Mapping the paths between self-construals, English language confidence, and adjustment. *International Journal of Intercultural Relations*, *30*(4), 487-506.

Yoon, G. Y. (2020). *Item performance in context: Differential item functioning between pilot and formal administration of the Norwegian language test* (Master's thesis).

Zieky, M. J. (2016). Fairness in test design and development. *Fairness in Educational Assessment and Measurement,* 9-32.

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*(3), 1-25.

Biodata

Mehri Jamalzadeh is a lecturer in the department of foreign languages, at Isfahan Azad University, Isfahan, Iran. She holds a Ph.D. in English Language Teaching (ELT) with a teaching experience of 5 years at university level. Her research interests are in the areas of corpus research, CALL, cultural and translation studies.
Email: *m82jamalzadeh@yahoo.com*

Dr. Ahmad Reza Lotfi is an assistant professor of English Language. Received his Ph.D. in English Language Teaching (ELT) from Ph.D. Research Centre of Islamic Azad University in Tehran. He completed my doctoral dissertation entitled On the Significance of Negative Evidence in Second-Language Learning under the supervision of Dr. A. Miremadi.
Email: *lotfi.ahmdrzlotfi@gmail.com*

Dr. Masoud Rostami is an assistant professor in the Department of Languages and Literature, at Yazd University, Yazd, Iran. He holds a Ph.D. in English language and literature, with a teaching experience of more than 15 years at university level. His research interests are in the areas of the philosophy of language and literature, cultural and translation studies.
Email: *mrostami@yazd.ac.ir*