Curriculum Research

Volume 6, Issue 3
Sept. 20, 2025

Fairness in High-stakes Testing: Analyzing Differential Item Functioning (DIF) by Gender, School type, and Ethnicity in Iran's National University Entrance Exam for Foreign Languages

Article Type:

Original Research

Authors:

Sanaz Behboudi Nehzomi[1]
ORCiD: 0009-0002-9591-2602

Masood Siyyari[2]
ORCiD: 0000-0002-6273-2739

Gholam-Reza Abbasian[3]
ORCiD: 0000-0003-1507-1736

Article History:

Received: 2025.05.27

Accepted: 2025.09.08

Published: 2025.09.20

Abstract

Numerous experts have underscored the need of fairness in National Entrance Examination items. This study examines whether examinees' performance on items of the National University Entrance Exam for Foreign Languages (NUEEFL), known as “Konkour,” varies based on background, specifically gender, school type, and ethnicity, rather than language proficiency, as the detection of differential item functioning (DIF) may enhance the fairness of high-stakes tests. The research employed a quantitative non-experimental, cross-sectional design. The participants included 200 male and female students, who were chosen randomly from students studying at Islamic Azad University, Science and Research branch in Tehran, Iran. The instruments consisted of a mock NUEEFL test and a researcher-made questionnaire. Upon taking the participants’ consent, the researcher took the mock version of NUEEFL. Next, the participants were asked to answer the questionnaire about their demographic information, including their gender, school type, and ethnicity. A three-phase DIF analysis was conducted to explore examinees' performance across these demographic variables. The results indicated that school type exhibited the most significant DIF, particularly in grammar and cloze assessments, whereas gender DIF was mostly seen in grammar and language function. Moreover, ethnically differential item functioning was significant in vocabulary and cloze assessments. Furthermore, reading comprehension was mostly impartial, with the exception of school type. The results underscore the need for test developers to consider demographic factors to ensure fairness and validity in high-stakes testing contexts.

Key Words: Differential Item Functioning, Ethnicity, Gender, School Type, Testing Fairness

1. Department of Language Teaching, SR.C., Islamic Azad University, Tehran, Iran. Email: sanaz.behboudi@srbiau.ac.ir
2. Department of Language Teaching, SR.C., Islamic Azad University, Tehran, Iran (Corresponding Author). Email: m.siyyari@srbiau.ac.ir
3. Department of English Language and Literature, Imam Ali University, Tehran, Iran. Email: gabbasian@gmail.com

1. Introduction

Student assessments have transformed from their original purpose of gauging job eligibility into a means of holding schools, instructors, and students accountable, as well as allocating educational resources (Yang et al., 2019). The aim of administering exams is to determine how well learners have learnt a given topic over a specified period of time to fulfill predetermined goals (Yang et al., 2019). For most students, it marks a major turning point in their scholastic path, opening doors to greater opportunities for growth. Testing was crucial in China about 200 BC for deciding who could hold public service jobs. It was common practice in Italy to grade educators by the end of the fourteenth century on the basis of their students' exam scores. These days, test scores are used for a lot of different things by politicians in different countries. First, they help with accountability (for students, teachers, and schools), second, they show where limited funds should go, and third, they inform political and government decisions that try to improve education (Russell et al., 2009).

The history of research on bias in testing has been contentious, particularly since Jensen's seminal work, "Bias in Mental Testing" (1980). Bias, in its statistical and ethical dimensions, remains a critical concern in educational assessments. A student's real knowledge or skills may not be accurately reflected by certain parts of the exam, leading to inaccurate or misleading judgments due to bias in testing. French (2020) points out that the test's psychometric qualities or general design might be the source of bias. Differences in prediction accuracy or the criteria used to pick the test are two examples of external variables that could introduce bias. According to French, bias is an inherent flaw in the measuring procedure that might have varying impacts on different groups' test results. Unfair results in student assessments may also result from assessment bias, which can have a disproportionate effect on pupils according to personal factors such as ethnicity, gender, socioeconomic level, religion, and origin (Kim & Zabelina, 2015).

As scholars, like Chalhoub-Deville (2015) and Kunnan (2018) argue, fairness in testing is paramount for social justice, necessitating valid and equitable measurement tools. Furthermore, a test must be equitable for diverse test takers. Specifically, it should not exhibit bias against the characteristics of the examinees (e.g., gender, race). Addressing this issue necessitates a statistical methodology for test analysis that can initially determine whether test items function differentially among various groups and ultimately identify the sources of this variance (Geranpayeh & Kunnan, 2007). One proposed method for this purpose is Differential Item Functioning (DIF). DIF arises when examinees with equivalent capability levels from distinct groups have differing probabilities of affirming an item (Mazor et al., 1998).

In Iran, the official name for the Iranian National University Exam is the “Konkour” examination. It is perhaps an altered version of the French word "concours," which means "to

source, screen, and select" and describes a variety of human resource management initiatives. Languages other than English are among the three main areas in which Konkour is offered. According to Razmjo (2006), this national English proficiency exam was first given to all citizens in 2002. In order to administer the exams, the National Organization of Educational Testing (NOET) collaborates with the major institutions in every city in the country. People who want to attend public/state universities that do not charge tuition have to pass this very tough test (Hosseini, 2007). The exam evaluates not just reading comprehension and vocabulary but also linguistic functions, the capacity to complete multiple-choice questions, and understanding of English syntax and structure. All of the materials included on the exam are taken from the textbooks that are used as primary resources for education in Iran. This assessment is used to gauge the students' proficiency and is known as a norm-referenced exam. Those who take the exam often are fluent Farsi speakers who visit designated testing locations (Khodi et al., 2021).

The importance assigned to Konkour generates concern in both candidates and their parents, due to its capacity to significantly influence their socio-economic standing (Parviz, 2023). A number of scholars have shown that Konkour remains a significant barrier to attending higher education institutions and is a substantial challenge to attaining educational parity (Ghorbani, 2012; Kamyab, 2007, 2008; Safari, 2016; Safari & Rashidi, 2018).

This study is significant as, in the Iranian context, it is assumed that most language tests in high stakes are not fair because they do not have validity (Safari, 2016). Due to the fact that the Konkoor determines examinees' future in terms of their study and career, as well as their personal life, it must be free from any kind of bias, and treats all examinees fairly (Khodi et al., 2021). Zumbo (1999) asserts that the concept of item bias has significant implications for policy, administration, and teaching settings. Consequently, bias may result in systemic mistakes that skew the conclusions drawn in the categorization and selection of pupils. This indicates that test-takers with comparable understanding of the test content should get equal results on individual questions, irrespective of their affiliation with other groups, including gender, culture, ethnicity, or race (Weijters et al., 2013). The concept of group is central to the definitions of bias (Davis, 2013), and this concept could be studied also in relation to other groups, such as social class, age, religion, or any other sociodemographic characteristic of the learners. As a result, this study was an attempt to investigate whether examinees' performances on test items differ due to their background, including gender, school type, and ethnicity rather than their language proficiency.

2. Review of the Related Literature

Fairness is seen as the core principle of classroom assessment notions and quality assessment techniques (Baniasadi et al., 2022; Rezai et al., 2021), consistently highlighted as a vital quality and a key trait of assessment-literate instructors (DeLuca et al., 2016). Mislevy (2018) posited fairness as a logical foundation for accommodating the interests and past knowledge of test-takers. He contendwd that comprehending job performance requires consideration of the individual, their past, the tasks, the environment, and the contextual factors.

The prospect of learning often prioritizes equity above evaluation and is closely linked to education (Rasooli et al., 2018). It comprises exposure to test material or more broadly refers to the alignment between curriculum and assessment (Tierney, 2016). Furthermore, it encompasses many educational possibilities tailored to learners' distinct learning styles, abilities, and exceptionalities (Lantolf & Poehner, 2013). An opportunity to showcase learning involves offering diverse, equitable, and meaningful chances for learners to display their knowledge (Tierney, 2016). Mauldin (2009) experimentally showed that evaluation is equitable when numerous assessment chances are provided, hence supporting the idea of fairness in assessment.

The non-existence of bias in a test is frequently employed as a component of the test fairness framework when data about the outcomes is gathered (Bachman, 2005). Test biases encompass: a) Offensive content or language, which pertains to material that is derogatory to examinees from diverse backgrounds, including stereotypes of group members and explicit or implicit slurs or insults related to gender, race and ethnicity, religion, age, language, national origin, and sexual orientation. b) Unjust penalization is influenced by the examinee's history, which pertains to material that may lead to inequitable penalties based on a test taker's group affiliation, including but not limited to gender, race, ethnicity, religion, age, language, country origin, and sexual orientation. c) Disparate influence and standard setting, which pertains to varying performances and following consequences among examinees from distinct group affiliations. To ensure fairness and validity in assessment, potential group disparities linked to key test-taker attributes, including gender, race, ethnicity, religion, age, language, national origin, and sexual orientation, should be rigorously investigated through Differential Item Functioning (DIF) and Differential Test Functioning (DTF) analyses at both the item and subtest levels (Zumbo, 1999). A differential validity study should be performed to see whether a test predicts success more effectively for one group compared to another. Test results should be evaluated about the criteria measure and selection judgments in the context of standard-setting. Test creators and consumers must be assured that valid measures and statistically robust, unbiased selection methods are used (Bachman, 2005). These studies

should educate test makers and users that group differences correlate with the tested abilities and that irrelevant constructs should not be included.

Khodi (2020) performed a generalizability research including 5000 Konkur examinees, demonstrating that 86% of the overall variation is due to individual differences, indicating a high degree of test reliability. The interaction between individuals' fields of study and the common questions in the test sections led to an error of around 1.5%. The national entrance examination demonstrates impartiality towards persons from diverse educational backgrounds. Also, Khodi et al. (2021) examined the admission examination for Iranian universities, referred to as "Konkour". Given the significance of this high-stakes examination, which may have social and long-term ramifications for the participants, they assessed the test and its psychometric properties. Their results suggested that the test offers a constrained context for assessing participants' "knowledge of language" rather than their "knowledge about language." Consequently, the dimensionality and validity of the assessment remain contentious.

In their study on graduate students' perceptions of assessment fairness, Darabi et al. (2022) found that "equity and interactional fairness" were the most important variables in deciding whether an assessment was fair or unfair. By interviewing 27 seasoned high school teachers, Rasooli et al. (2022) sought to understand what factors influence teachers' perspectives on equality in the classroom. Individual mechanisms, societal mechanisms, and the dialectical links between the two, were the three main topics that affected how instructors perceived assessment fairness. The research showed that teachers' views and behaviors on assessment fairness are shaped by a complex web of factors, including their own beliefs and experiences as well as the social, institutional, and classroom settings in which they work.

Dadvar and Tabatabaee-Yazdi (2023) examined the correlation between Iranian EFL learners' views of equity in assessment and their cognitive test anxiety. The research used a quantitative methodology and included a total of 201 participants. A convenience and random sample approach was used, using a fair evaluation and a cognitive test anxiety questionnaire. The study's findings, derived from Pearson-moment correlation and regression analysis, indicated no significant association between Iranian EFL learners' judgments of fairness in classroom assessment and their cognitive test anxiety. Besides, no substantial difference between male and female learners' sense of fairness in classroom evaluation and their cognitive test anxiety was shown.

Educators of English as a foreign language (EFL) have their perceptions of fairness in evaluation (Tofighi & Ahmadi Safa, 2023). These two researchers first aimed to develop and assess a Classroom Assessment Fairness Scale (CAFS) as part of a whole methodology. A

total of 120 EFL teachers in Iran were given the validated scale. According to the results, EFL educators had a deep comprehension of what constitutes a fair assessment strategy for the classroom. Equity in classroom assessment techniques was identified as critically important by their recognition of the importance of learning opportunities, the capacity to show knowledge, a positive classroom atmosphere, the no-harm principle, and openness. Score pollution was not, however, guaranteed to have the same level of familiarity or awareness. Further, there were statistically significant variations in how EFL teachers perceived the fairness of classroom assessment based on instructors' gender, educational background, years of experience teaching, and learning environment.

Parviz (2023) performed an exhaustive analysis of the previous developments and current status of Konkour. A systematic-narrative hybrid literature review (SNHLR) was used to meet the study's aims. Seventy-four peer-reviewed research papers, both national and international, were meticulously selected and analyzed to identify significant themes, patterns, gaps, and trends regarding the ramifications of Konkour. The findings revealed that Iranian university applicants had faced many negative consequences, including economic, psychological, and educational impacts owing to the Konkour.

Due to the importance of test fairness, numerous studies have been carried out and various models have been proposed in the international and Iranian contexts (e.g., Dadvar & Tabatabaee-Yazdi, 2023; Haertel & Herman, 2005; McNamara & Roever, 2006; Shohamy & Eldar, 2000; Tofighi & Ahmadi Safa, 2023). However, the studies conducted so far do not yield a compelling account of the fairness associated with the Konkoor examination in the Iranian EFL context. They only propose the general constructs of fairness without going into details of the issue (e.g., Kodi et al., 2021). Thus, the present research aimed to provide a quantified and objective account of fairness in the Konkoor examination to fill the gap in the literature. Therefore, the overarching research question was as follows:

- Do examinees perform differently on any test items as a result of their background (e.g., gender, school type, or ethnicity) rather than their language proficiency?

3. Method

3.1. Design

The design of the present study was a quantitative, non-experimental, cross-sectional research design with a focus on DIF analysis (Ary et al., 2018)**.** Data were collected at a single point in time (after administering the mock NUEEFL test and the demographic questionnaire). It did not involve experimental manipulation but rather examined naturally occurring group

differences in test performance.

3.2. Participants

The participants of the present study included B.A. Konkoor candidates, including 200 students of both genders (100 male and 100 female) who had taken National University Entrance Exam for Foreign Languages (NUEEFL) known as “Konkour”. They were selected randomly from university students from different regions studying at Islamic Azad University, Science and Research branch in Tehran, Iran. The first language of the participants was Persian, and their age ranged between 18-20 years old. Prior to the study, written consent regarding the participation in this study was obtained from all the participants.

3.3. Instruments

3.3.1. Konkour Test. A mock Konkour was administered to the participants. The purpose of this mock Konkour test was to evaluate the test in terms of its differential item functioning. The exam was structured into six components, each reflecting the content of the textbooks provided to high school students. The test had 70 items as follows:

Grammar Test: This section comprised ten questions about English grammar. The questions were formatted as unfinished sentences that students had to complete by selecting from alternatives, which included phrases, nouns, prepositions, or verbs. The succession of questions was not governed by a predetermined rule; instead, it was arranged randomly for each participant, rather than uniformly across all participants. Occasionally, two grammatical rules were conflated inside a single question, making it very complicated for students to discern the underlying concept and identify the proper answer.

Vocabulary Test: This section included 15 questions formatted as incomplete sentences. The students had to choose the best alternative for completing the meaning of the sentences. The right answer for the participants was already communicated in the classroom environment, however the distractors consisted of unfamiliar vocabulary for them. The parts of speech could vary among questions, although efforts were made to maintain consistency across the possibilities of each individual item to reduce the likelihood of random guessing by participants.

Sentence Structure: In this part, there were five questions; each item of the question presented a sentence and the participants had to select the item in which there was no grammatical mistake based on the stem of the question. The sentences were relatively long, often containing complex structures, and errors could occur in any part of the phrase.

Language Functions: This section was consisted of various conversations and comprised ten questions. Participants had to complete the dialogues with the most appropriate responses from the provided alternatives. The accurate response should operate as a complement to the

interaction occurring between both parties in the conversation.

Cloze Test: In this section, participants were required to read a text containing 15 blanks, almost positioned at regular intervals, such as every ten words, and choose the choice that most effectively completed each sentence. The presentation of blanks inside a single text could result in misunderstandings or errors in identifying the right answer for one blank, perhaps leading participants to pick inappropriate options for subsequent blanks.

Reading Comprehension: Each exam encompassed three reading comprehension passages, each ranging from 350 to 500 words. The passages included a diverse array of subjects, such as scientific and social themes. Each text included five multiple-choice questions about the topic, word meanings, and sentence interpretations.

3.3.2. Demographic Questionnaire. A researcher-designed questionnaire was employed in the present study to collect participants' demographic information, including gender, school type, and ethnicity.

3.4. Procedure

Participants were randomly selected from among B.A. Konkour candidates, consisting of 200 students of both genders who had taken the National University Entrance Examination of Foreign Languages (NUEEFL), commonly known as the Konkour, at the Islamic Azad University, Science and Research Branch. The researcher explained the study's objectives to the participants, obtained their informed consent, and then administered a mock version of the NUEEFL. The test was conducted in a single 100-minute session. Following the exam, participants completed a demographic questionnaire providing information on gender, school type, and ethnicity. The participants differed in age and field of study. While all were proficient in Persian as the medium of instruction, their native languages were not systematically collected and therefore are not reported. These data were subsequently analyzed using Differential Item Functioning (DIF) to address the research question, with examinees' performance on the NUEEFL examined in relation to gender, school type, and ethnicity.

3.5 Data Analysis

The collected data were analyzed using **Differential Item Functioning (DIF)** to examine the fairness of the NUEEFL across gender, school type, and ethnicity. After data screening and descriptive statistics, DIF analysis was used to identify whether test items functioned differently for subgroups of examinees who were matched on overall ability. This approach allowed the study to detect potential sources of bias and evaluate the extent to which item performance reflected differences in construct-relevant ability rather than group membership.

4. Results

4.1. Demographic Characteristics of the Participants

Table 1 displays the demographic characteristics of the participants. With respect to ethnicity, Fars (34.2%) and Turk (32.3%) students constituted the largest groups, while Kurd (16.9%) and Gilaki (16.5%) students accounted for smaller but still substantial portions of the sample. The gender distribution reveals a clear imbalance, with females representing nearly three-quarters of the participants (74.8%) and males comprising only one-quarter (25.2%). Regarding school type, a majority of the examinees attended private schools (56.0%), whereas 44.0% were enrolled in public schools.

Table 1.

Frequencies and Percentages of Demographic Information

		Frequency	Percent
Ethnicity	Fars	91	34.2
	Gilaki	44	16.5
	Kurd	45	16.9
	Turk	86	32.3
	Total	266	100.0
Gender	Female	199	74.8
	Male	67	25.2
	Total	266	100.0
School Type	Private	149	56.0
	Public	117	44.0
	Total	266	100.0

4.2. Differential Item Functioning on Gender

Table 2 displays the results of the DIF for the grammar sub-section of NUEEFL. The results indicated that there were significant DIF's on first (χ^2 = 16.95, p = .002), fourth (χ^2 = 10.26, p = .020), and seventh items (χ^2 = 13.35, p = .006). Before discussing the results, it should be noted that two sets of p-values were computed. The last column included the p-values adjusted for multiple comparisons made to reduce the inflated error rate.

Table 2.

Differential Item Functioning Grammar by Gender

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item1	16.952	0.000	0.002	**
Item2	1.401	0.497	0.621	
Item3	0.336	0.845	0.845	
Item4	10.264	0.006	0.020	*
Item5	4.850	0.089	0.221	
Item6	1.866	0.393	0.562	
Item7	13.352	0.001	0.006	**
Item8	2.074	0.355	0.562	

Item9	1.104	0.576	0.640
Item10	3.326	0.190	0.379

Table 3 displays the results of DIF which compare male and female groups' performance on 15 items of the vocabulary test. The results showed no significant DIF's for the two groups on vocabulary test.

Table 3.

Differential Item Functioning Vocabulary by Gender

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item11	4.435	0.109	0.611	
Item12	3.630	0.163	0.611	
Item13	0.954	0.621	0.776	
Item14	1.335	0.513	0.776	
Item15	2.139	0.343	0.735	
Item16	1.720	0.423	0.776	
Item17	2.603	0.272	0.680	
Item18	0.253	0.881	0.881	
Item19	0.960	0.619	0.776	
Item20	0.428	0.807	0.881	
Item21	4.096	0.129	0.611	
Item22	0.299	0.861	0.881	
Item23	3.686	0.158	0.611	
Item24	2.950	0.229	0.680	
Item25	1.177	0.555	0.776	

Table 4 displays the results of DIF which compare male and female groups' performance on five items of the sentence structure test. The results showed no significant DIF's for the two groups on sentence structure test.

Table 4.

Differential Item Functioning Sentence Structure by Gender

	Chi-square	**p-value**	**Adjusted p-value**	**Significant DIF**
Item26	0.330	0.848	0.848	
Item27	1.337	0.513	0.848	
Item28	3.400	0.183	0.848	
Item29	1.142	0.565	0.848	
Item30	0.609	0.738	0.848	

Table 5 displays the results of the DIF for the language function sub-section of NUEEFL. The results indicated that there were significant DIF's on item 34 ($\chi^2 = 15.95$, $p = .002$), and item 39 ($\chi^2 = 12.00$, $p = .012$).

Table 5.

Differential Item Functioning Language Function by Gender

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item31	1.421	0.492	0.614	
Item32	0.984	0.611	0.679	
Item33	7.964	0.019	0.062	
Item34	15.954	0.000	0.003	**
Item35	0.318	0.853	0.853	
Item36	3.560	0.169	0.281	
Item37	4.474	0.107	0.214	
Item38	2.304	0.316	0.451	
Item39	12.004	0.003	0.012	*
Item40	4.597	0.100	0.214	

Table 6 displays the results of DIF which compare male and female groups' performance on 15 items of the cloze test. The results showed no significant DIF's for the two groups on cloze test.

Table 6.

Differential Item Functioning Cloze Test by Gender

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item41	0.549	0.760	0.966	
Item42	0.069	0.966	0.966	
Item43	0.389	0.823	0.966	
Item44	0.422	0.810	0.966	
Item45	1.315	0.518	0.966	
Item46	1.964	0.375	0.966	
Item47	1.792	0.408	0.966	
Item48	0.326	0.850	0.966	
Item49	3.047	0.218	0.966	
Item50	6.168	0.046	0.687	
Item51	0.256	0.880	0.966	
Item52	1.059	0.589	0.966	
Item53	3.459	0.177	0.966	
Item54	2.321	0.313	0.966	
Item55	0.134	0.935	0.966	

Finally, Table 7 displays the results of DIF which compare male and female groups' performance on 15 items of the reading comprehension test. The results showed no significant DIF's for the two groups on reading comprehension test.

Table 7.

Differential Item Functioning Reading Comprehension by Gender

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item56	0.004	0.998	0.998	
Item57	0.643	0.725	0.998	
Item58	0.066	0.968	0.998	
Item59	0.704	0.703	0.998	
Item60	2.937	0.230	0.998	
Item61	0.032	0.984	0.998	
Item62	1.197	0.550	0.998	
Item63	0.209	0.901	0.998	
Item64	2.594	0.273	0.998	
Item65	1.461	0.482	0.998	
Item66	1.179	0.555	0.998	
Item67	0.313	0.855	0.998	
Item68	1.498	0.473	0.998	
Item69	2.711	0.258	0.998	
Item70	5.362	0.069	0.998	

4.2. Differential Item Functioning on School Type

Table 8 displays the results of the DIF for the grammar sub-section of NUEEFL by school type. The results indicated that there were significant DIF's on second (χ^2 = 7.60, p = .045), third (χ^2 = 32.90, p = .000), fourth (χ^2 = 13.78, p = .003), fifth (χ^2 = 531.02, p = .000), and tenth (χ^2 = 652.96, p = .000) items. Before discussing the results, it should be noted that the private schools were labeled as focal.

Table 8.

Differential Item Functioning Grammar by School Type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item1	0.226	0.893	0.893	
Item2	7.609	0.022	0.045	*
Item3	32.909	0.000	0.000	***
Item4	13.782	0.001	0.003	**
Item5	531.025	0.000	0.000	***
Item6	2.509	0.285	0.357	
Item7	3.297	0.192	0.275	
Item8	1.889	0.389	0.432	
Item9	5.346	0.069	0.115	
Item10	652.961	0.000	0.000	***

Table 9 displays the results of DIF which compare public and private schools' performance on 15 items of the vocabulary test. The results showed there were significant

DIF's on items 14 (χ^2 = 17.65, p = .002), and 23 (χ^2 = 10.04, p = .049).

Table 9.

Differential Item Functioning Vocabulary by School Type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item11	4.885	0.087	0.145	
Item12	7.208	0.027	0.068	
Item13	2.412	0.299	0.321	
Item14	17.653	0.000	0.002	**
Item15	7.775	0.021	0.068	
Item16	2.439	0.296	0.321	
Item17	0.267	0.875	0.875	
Item18	5.551	0.062	0.117	
Item19	2.498	0.287	0.321	
Item20	3.577	0.167	0.251	
Item21	6.875	0.032	0.069	
Item22	7.434	0.024	0.068	
Item23	10.044	0.007	0.049	*
Item24	9.005	0.011	0.055	
Item25	2.980	0.225	0.307	

Table 10 displays the results of DIF which compare public and private schools' performance on five items of the sentence structure test. The results showed no significant DIF's for the two groups on sentence structure test.

Table 10.

Differential Item Functioning Sentence Structure by School Type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item26	3.676	0.159	0.746	
Item27	1.608	0.447	0.746	
Item28	0.574	0.751	0.938	
Item29	0.023	0.989	0.989	
Item30	2.259	0.323	0.746	

Table 11 displays the results of the DIF for the language function sub-section of NUEEFL. The results indicated that there were not any significant DIF's for ten items of language function.

Table 11.

Differential Item Functioning Language Function by School Type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item31	0.123	0.940	0.940	
Item32	1.463	0.481	0.688	

Item33	5.981	0.050	0.181
Item34	5.715	0.057	0.181
Item35	0.602	0.740	0.910
Item36	5.249	0.073	0.181
Item37	1.880	0.391	0.688
Item38	1.536	0.464	0.688
Item39	0.399	0.819	0.910
Item40	10.442	0.005	0.054

Table 12 displays the results of DIF which compare public and private schools' performance on 15 items of the cloze test. The results showed there were significant DIF's on 11 items out of the 15 items of the cloze test; item 42 (χ^2 = 12.28, p = .007), item 45 (χ^2 = 8.40, p = .025), item 47 (χ^2 = 7.96, p = .029), item 48 (χ^2 = 8.40, p = .028), item 49 (χ^2 = 17.92, p = .001), item 50 (χ^2 = 18.18, p = .001), item 51 (χ^2 = 6.88, p = .044), item 52 (χ^2 = 27.17, p = .000), item 53 (χ^2 = 10.44, p = .014), item 54 (χ^2 = 7.86, p = .029), and item 55 (χ^2 = 22.29, p = .000).

Table 12.

Differential Item Functioning Cloze Test by School type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item41	4.704	0.095	0.102	
Item42	12.281	0.002	0.007	**
Item43	5.941	0.051	0.064	
Item44	1.108	0.575	0.575	
Item45	8.402	0.015	0.028	*
Item46	5.048	0.080	0.093	
Item47	7.967	0.019	0.029	*
Item48	8.408	0.015	0.028	*
Item49	17.923	0.000	0.001	***
Item50	18.189	0.000	0.001	***
Item51	6.880	0.032	0.044	*
Item52	27.176	0.000	0.000	***
Item53	10.441	0.005	0.014	*
Item54	7.868	0.020	0.029	*
Item55	22.299	0.000	0.000	***

Finally, Table 13 displays the results of DIF which compare public and private schools' performance on 15 items of the reading comprehension test. The results showed there were two significant DIF's on items 59 (χ^2 = 24.71, p = .000), and 62 (χ^2 = 10.46, p = .040).

Table 13.

Differential Item Functioning Reading Comprehension by School type

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item56	2.405	0.300	0.349	
Item57	3.591	0.166	0.277	
Item58	8.909	0.012	0.058	
Item59	24.710	0.000	0.000	***
Item60	0.093	0.955	0.974	
Item61	6.690	0.035	0.078	
Item62	10.463	0.005	0.040	*
Item63	2.792	0.248	0.349	
Item64	6.621	0.037	0.078	
Item65	5.196	0.074	0.140	
Item66	2.393	0.302	0.349	
Item67	7.615	0.022	0.078	
Item68	0.053	0.974	0.974	
Item69	7.194	0.027	0.078	
Item70	2.606	0.272	0.349	

4.3. Differential Item Functioning on Ethnicity

Table 14 displays the results of the DIF for the grammar sub-section of NUEEFL by ethnicity. The results indicated that there were significant DIF's on first ($\chi^2 = 10.32$, $p = .029$), and tenth ($\chi^2 = 17.33$, $p = .002$) items.

Table 14.

Differential Item Functioning Grammar by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item1	10.320	0.006	0.029	*
Item2	0.825	0.662	0.769	
Item3	0.022	0.989	0.989	
Item4	0.795	0.672	0.769	
Item5	0.736	0.692	0.769	
Item6	2.351	0.309	0.617	
Item7	5.181	0.075	0.250	
Item8	1.684	0.431	0.718	
Item9	3.279	0.194	0.485	
Item10	17.335	0.000	0.002	**

Table 15 displays the results of DIF which compare ethnicity groups' performance on 15 items of the vocabulary test. The results showed there were significant DIF's on items 14 (χ^2

= 9.27, p = .036), 15 (χ^2 = 13.36, p = .009), 19 (χ^2 = 31.74, p = .000), and 21 (χ^2 = 11.65, p = .015).

Table 15.

Differential Item Functioning Vocabulary by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item11	6.655	0.036	0.067	
Item12	4.726	0.094	0.157	
Item13	1.325	0.516	0.552	
Item14	9.271	0.010	0.036	*
Item15	13.367	0.001	0.009	**
Item16	0.250	0.883	0.883	
Item17	3.319	0.190	0.259	
Item18	3.519	0.172	0.258	
Item19	31.749	0.000	0.000	***
Item20	2.171	0.338	0.422	
Item21	11.659	0.003	0.015	*
Item22	7.434	0.024	0.067	
Item23	6.797	0.033	0.067	
Item24	6.872	0.032	0.067	
Item25	1.691	0.430	0.496	

Table 16 displays the results of DIF which compare four ethnicity groups' performance on five items of the sentence structure test. The results showed a single significant DIF on item 30; i.e. (χ^2 = 15.58, p = .002).

Table 16.

Differential Item Functioning Sentence Structure by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item26	0.418	0.811	0.811	
Item27	4.086	0.130	0.324	
Item28	2.588	0.274	0.457	
Item29	1.803	0.406	0.507	
Item30	15.589	0.000	0.002	**

Table 17 displays the results of the DIF for the language function sub-section of NUEEFL. The results indicated that there were not any significant DIF's on ten items of language function.

Table 17.

Differential Item Functioning Language Function by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item31	0.807	0.668	0.833	
Item32	0.365	0.833	0.833	
Item33	3.659	0.161	0.321	
Item34	6.412	0.041	0.135	
Item35	2.727	0.256	0.365	
Item36	3.853	0.146	0.321	
Item37	7.641	0.022	0.112	
Item38	2.851	0.240	0.365	
Item39	0.413	0.814	0.833	
Item40	7.601	0.022	0.112	

Table 18 displays the results of DIF which compare ethnicity groups' performance on 15 items of the cloze test. The results showed there except for the first two items; item 43 to 55 showed significant DIF across three ethnicity groups.

Table 18.

Differential Item Functioning Cloze Test by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item41	0.000	1.000	1.000	
Item42	0.000	1.000	1.000	
Item43	42.857	0.000	0.000	***
Item44	22.118	0.000	0.000	***
Item45	49.738	0.000	0.000	***
Item46	49.566	0.000	0.000	***
Item47	17.922	0.000	0.000	***
Item48	61.123	0.000	0.000	***
Item49	5.949	0.015	0.017	*
Item50	31.741	0.000	0.000	***
Item51	23.346	0.000	0.000	***
Item52	38.598	0.000	0.000	***
Item53	130.979	0.000	0.000	***
Item54	80.333	0.000	0.000	***
Item55	64.650	0.000	0.000	***

Finally, Table 19 displays the results of DIF which compare ethnicity groups' performance on 15 items of the reading comprehension test. The results showed there were not any significant DIF's on 15 items of reading comprehension section of NUEEFL.

Table 19.

Differential Item Functioning Reading Comprehension by Ethnicity

	Chi-square	p-value	Adjusted p-value	Significant DIF
Item56	3.840	0.147	0.459	
Item57	3.391	0.184	0.459	
Item58	2.352	0.309	0.490	
Item59	1.364	0.506	0.632	
Item60	4.467	0.107	0.459	
Item61	0.537	0.765	0.819	
Item62	2.238	0.327	0.490	
Item63	2.927	0.232	0.490	
Item64	1.184	0.553	0.638	
Item65	4.977	0.083	0.459	
Item66	6.773	0.034	0.459	
Item67	1.505	0.471	0.632	
Item68	0.004	0.998	0.998	
Item69	2.268	0.322	0.490	
Item70	3.453	0.178	0.459	

5. Discussion

The results indicated significant DIF related to gender, school type, and ethnicity in the NUEEFL, highlighting serious issues with fairness, construct validity, and possible bias in high-stakes language evaluation. The results indicated that school type exhibited the most significant DIF, particularly in grammar and cloze test assessments, whereas gender DIF was mostly seen in grammar and language function. Moreover, ethnically differential item functioning was significant in vocabulary and cloze test assessments, with item 19 (vocabulary) exhibiting the most significant bias. Furthermore, reading comprehension was mostly impartial, with the exception of school type. The findings suggest that the grammar and language function sections may include elements that preferentially benefit one gender, necessitating more examination. The cloze test and certain grammar components may disfavor learners from both public and private schools, prompting issues over fairness in test design. Moreover, the vocabulary and cloze test sections may include culturally or linguistically biased elements, thus influencing performance across different ethnic groups.

The most pervasive DIF was observed between public and private schools, particularly in grammar and cloze tests. Items such as grammar questions 5 and 10**,** exhibited extreme bias suggesting that students from different school backgrounds may interpret or respond to these items, differently. Private schools may emphasize certain grammatical structures or test-taking strategies not equally covered in public schools. Moreover, access to quality English

instruction, tutoring, or practice materials may vary significantly between school types. In addition, private school students may have more opportunities for immersive English exposure.

Significant DIF was found for male and female test-takers **in** grammar (items 1, 4, 7) and language function (items 34, 39)**.** Prior research suggests that males and females may employ different strategies in grammatical reasoning or pragmatic language use. Furthermore, certain topics or phrasing in these items may resonate differently across genders. Ethnicity-related DIF was most prominent in vocabulary (item 19) and the cloze test (items 43–55)**.** Some vocabulary items (e.g., item 19) may reflect concepts more familiar to certain ethnic groups. Moreover, cloze passages may contain structures more aligned with specific linguistic backgrounds.

Reading comprehension showed minimal DIF (only two items affected by school type), suggesting it may be the most robust section for fair assessment. Reading tasks assess skills less dependent on isolated grammar/vocabulary knowledge. This finding lend credence to those Gonzalez and Hinton (2018) who found that reading comprehension items had minimal DIF across various educational environments.

The findings of the current study revealed that the NUEEFL may fail to satisfy norms of fairness and dependability. The results are in line with other studies that have raised questions about the accuracy, fairness, and equity of high-stakes language tests in different settings (Alderson & Hamp-Lyons, 1996; Bachman, 1990; Shohamy & Eldar, 2002). In particular, aligning with Alderson and Hamp-Lyons (1996), the findings suggest that these tests may favor specific groups of learners, consequently challenging assertions of fairness. Bachman's (1990) framework of test usefulness aligns with the current study by illustrating that construct validity may be undermined when contextual and social factors, such as gender or ethnicity, disproportionately affect performance. In the same way, Shohamy and Eldar (2002) stressed the sociopolitical aspects of testing, showing how high-stakes exams may make things worse instead of giving everyone a fair chance. This is also true of the current findings.

In the Iranian context, the findings validate previous research that has emphasized pervasive inequities in the NUEEFL. For instance, Kamyab (2007, 2008) and Ghorbani (2012) recorded problems of construct underrepresentation and content imbalance, which are similar to the item-level biases found in this study. Safari (2016) and Safari and Rashidi (2018) highlighted the sociocultural and educational disadvantages encountered by students from public schools and rural regions, findings that correspond with the current study's evidence of differential item functioning across school types. Khodi et al. (2021) and Parviz (2023) have recently reported that there are still differences in access, performance, and fairness in the

NUEEFL based on gender and region. This is in line with what we found in the current analysis.

The results corroborate those of Khodi et al. (2021), who determined that although the test context and content align, the nature of the assessment is not a reliable predictor of participants' actual language proficiency owing to the absence of some essential language abilities in the evaluation. Furthermore, the results go against what Khodi (2020) found, which is that the national entrance exam does not discriminate against any group of people based on their level of education.

Despite the fact that the designers of Konkour took into mind the concept of linguistic competency, fully operationalizing it proved to be difficult owing to practical limits and test-related issues. This is reflected in the style and structure of the items. To be more specific, the productive and receptive abilities of speaking, writing, and listening were not included in the evaluation. This was primarily due to the fact that there were regional differences in availability to training materials and the logistical issues that were involved in delivering such components throughout the country. According to Khodi et al. (2021), their absence was also justified on the basis of eliminating subjectivity in scoring. This was due to the fact that it is difficult to ensure inter-rater reliability and fairness in large-scale testing. In light of this, Konkour has placed an emphasis on skills that can be objectively measured, which has resulted in an increase in reliability and practicality. However, this has come at the expense of construct representativeness. With this trade-off, some logistical and equitable problems are addressed; but, at the same time, the breadth of language ability that is being evaluated is narrowed, which increases the danger of creating graduates who have limited communicative competence. The current analysis validates the use of Differential Item Functioning (DIF) analyses to give empirical evidence of fairness. These design restrictions also heighten the need of assessing whether test items work fairly across subgroups, which is why the present study used these analyses.

6. Conclusion

The current research sought to investigate whether examinees' performance on the NUEEFL varied based on background, specifically, gender, school type, and ethnicity, rather than language competency. DIF analyses revealed that the majority of test items performed equitably across the groups, however several items had differential functioning, indicating possible sources of bias. These findings highlight the need of overseeing and enhancing high-stakes language tests to guarantee they accurately assess language competency and provide fair possibilities for all test-takers.

This study highlights the complex interplay of gender, school type, ethnicity and English

language proficiency. Addressing these factors through equitable testing practices and inclusive curricula can enhance language learning outcomes for all students. Future research should continue to explore these dynamics to inform more just and effective educational assessments. As educational environments become increasingly diverse, it is essential to develop and implement policies and practices that recognize and support this diversity, ensuring that all students have the opportunity to achieve high levels of language proficiency.

The study provides important insights into the fairness of the NUEEFL. These findings suggest that improvements are needed to ensure that the test is reliable, valid, and fair for all examinees, regardless of their gender, school type, or ethnicity. This study has also implications for policymakers, test developers, and educators who need to address these issues and ensure that the test meets international standards of fairness. The findings underscore the need for educational policies that promote fairness and inclusivity in testing. Test developers must consider the diverse backgrounds of test-takers to ensure equitable assessment practices. Educators should adopt culturally responsive teaching methods to support all students, particularly those from minority backgrounds. Policymakers should advocate for curricula that integrate multicultural perspectives and provide adequate resources for public schools to offer high-quality language education. Training programs for teachers should emphasize the importance of cultural competence and gender sensitivity in language instruction.

The present study offers several important implications for Konkour test designers too. It emphasizes meticulously analyzing test questions for gender, school type, and ethnicity fairness. The content should be checked to ensure that it does not bias or disadvantage certain groups. Second, the test constructors should provide a more complete picture of language abilities to better assess English proficiency. This should encompass information production and reception. This study shows that conventional test validation procedures require statistical methods, like DIF. This would enable empirical bias detection and correction. The test designers must be proactive about equality to boost Konkour's legitimacy, validity, and social acceptability as a high-stakes language exam. This may be achieved by matching item selection and scoring with global fairness norms.

The university entrance examination is clearly a high-stake test whose results have grave consequences for the test takers. Failure on the test cannot be easily ignored. This failure might be due to failure on a single item. Therefore, it is the responsibility of the test developers and users to make sure that each and every item included in the test is fair. Our results show that this is not the case. All in all, by addressing the issues presenting in the present study, the NUEEFL can better serve its purpose as a fair and valid measure of English proficiency for all test-takers, regardless of background.

This study proposes some solutions to enhance the fairness and validity of the NUEEFL. Test designers should use DIF analysis on a frequent basis to find and fix biased items. This will make sure that all things operate the same for all genders, school types, and ethnicities. Second, inclusive educational policies and curricula should provide suitable resources for public schools and minority groups. In addition, educators should employ culturally responsive teaching approaches and train other teachers about cultural competence, gender sensitivity, and fair language instruction so that they can fulfill the needs of all of their students. The test should be piloted in a real-world setting to make sure it is reliable, legitimate, and fair for all test-takers, regardless of their background.

This research encompassed several limitations that could have influenced the outcomes. Initially, the participants in the study might not accurately reflect the diverse population of NUEEFL test-takers in Iran, as the data collection was limited to a selected group of schools due to practical constraints. Secondly, obtaining comprehensive demographic and performance data was not always feasible, complicating the exploration of the interconnections among specific subgroups. Thirdly, a more comprehensive understanding of the factors contributing to unequal item functioning could have emerged if the research had incorporated additional qualitative perspectives from educators or test-takers alongside the quantitative test data. Finally, while the sample size was adequate for DIF analysis, it was insufficient for a thorough examination of less prevalent ethnic groupings, potentially limiting the statistical capacity to detect minor biases. To address these challenges and yield more reliable and broadly applicable findings, subsequent research should consider employing larger sample sizes, diverse data sources, and more extensive sampling methods.

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297. https://doi.org/10.1177/026553229601300304

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

Baniasadi, A., Salehi, K., Khodaie, E., Bagheri Noaparast, Kh., & Balal Izanloo, B. (2023). Fairness in classroom assessment: A systematic review. *Asia-Pacific Education Researcher, 32*(1), 91–109. https://doi.org/10.1007/s40299-021-00636-z

Chalhoub-Deville, M. (2015). Testing context: Test-taker characteristics. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 405–418). Routledge.

Dadvar, N., & Tabatabaee-Yazdi, M. (2023). EFL learners' perceptions of fairness in classroom assessment and their cognitive test anxiety. *Journal of Research in Techno-based Language Education, 3*(4). https://doi.org/10.22034/jrtle.2023.413067.1095

Darabi Bazvand, A., & Rasooli, A. (2022). Students' experiences of fairness in summative assessment: A study in a higher education context. *Studies in Educational Evaluation, 72*, 101118.

https://doi.org/10.1016/j.stueduc.2021.101118

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education, 34*(5–6), 576–591. https://doi.org/10.1080/01626620.2012.730347

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

French, B. F. (2020). Test bias. In *Encyclopedia of quality of life and well-being research* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-319-69909-7_3261-1

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the Certificate in Advanced English examination. *Language Assessment Quarterly, 4*(2), 190–222. https://doi.org/10.1080/15434300701375758

Ghorbani, M. R. (2012). Controversy over abolishing Iranian university entrance examination. *Asian Education and Development Studies, 1*(2), 139–152. https://doi.org/10.1108/20463161211240181

Ghorbani, V., & Kianifard, S. (2024). Exploring ways of assessing intercultural competence: Introducing the bimodal assessment model. *Curriculum Research, 5*(4), 27–48. https://doi.org/10.71703/cure.2024.1129165

Haertel, E., & Herman, J. (2005). Historical perspective on validity arguments for accountability testing (CSE Report 654). University of California.

Hosseini, S. M. H. (2007). ELT in higher education in Iran and India: A critical view. *Language in India, 7*, 1–11.

Kamyab, S. (2007). Flying brains: A challenge facing Iran today. *International Higher Education, 47*. https://doi.org/10.6017/ihe.2007.47.7955

Khodi, A. (2020). *An appraisal of validity and dimensionality of B.A. Iranian University Entrance Examination* (Unpublished doctoral dissertation). University of Tehran.

Khodi, A., Alavi, S. M., & Karami, H. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Language Testing in Asia, 11*(14), 1–10. https://doi.org/10.1186/s40468-021-00125-6

Kim, K. H., & Zabelina, D. (2015). Cultural bias in assessment: Can creativity assessment help? *International Journal of Critical Pedagogy, 6*(2), 130–148.

Kunnan, A. J. (2018). *Fairness and justice in language assessment*. Cambridge University Press. https://doi.org/10.1017/9781139649484

Mauldin, R. K. (2009). Gendered perceptions of learning and fairness when choice between exam types is offered. *Active Learning in Higher Education, 10*(3), 253–264. https://doi.org/10.1177/1469787409343191

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357–367. https://doi.org/10.1177/014662169802200404

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. https://doi.org/10.4324/9781315708302

Parviz, M. (2023). Reflecting on the consequences of the Iranian university entrance examination: A systematic-narrative hybrid literature review. *Discover Education, 2*(22). https://doi.org/10.1007/s44217-023-00046-x

Rasooli, A., Rasegh, A., Zandi, H., & Firoozi, T. (2022). Teachers' conceptions of fairness in classroom assessment: An empirical study. *Journal of Teacher Education, 74*(3), 260–273. https://doi.org/10.1177/00224871221130742

Rasooli, A., Zandi, H., & DeLuca, C. (2018). Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond. *Studies in Educational*

Evaluation, 56, 164–181. https://doi.org/10.1016/j.stueduc.2017.12.008

Razmjo, S. A. (2006). Content analysis of specific questions of the English language test group of the national entrance exam of the country's universities. *Shiraz University Journal of Social Sciences and Humanities, 1*(46), 465–480.

Rezai, A., Alibakhshi, G., Farokhipour, S., & Miri, M. (2021). A phenomenographic study on language assessment literacy: Hearing from Iranian university teachers. *Language Testing in Asia, 11*, 1–18. https://doi.org/10.1186/s40468-021-00142-5

Russell, M., Madaus, G., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. LAP.

Safari, P. (2016). Reconsideration of language assessment is a MUST for democratic testing in the educational system of Iran. *Interchange, 47*(3), 267–296. https://doi.org/10.1007/s10780-016-9282-4

Safari, P., & Rashidi, N. (2018). Democratic assessment as scales of justice: The case of three Iranian high-stakes tests. *Policy Studies, 39*(2), 127–144. https://doi.org/10.1080/01442872.2018.1428748

Shohamy, E., & Eldar, S. (2002). High stakes exams and washback: The case of the Bagrut in Israel. *Assessment in Education, 9*(3), 307–333. https://doi.org/10.1080/0969594022000027646

Tierney, R. (2016). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1–6). Springer. https://doi.org/10.1007/978-981-287-532-7_171-1

Tofighi, S., & Ahmadi Safa, M. (2023). Fairness in classroom language assessment from EFL teachers' perspective. *Teaching English as a Second Language Quarterly, 42*(2), 81–110. https://doi.org/10.22099/tesl.2023.46825.3173

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320–334. https://doi.org/10.1037/a0032125

Xu, Y., & Brown, G. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Yang, B. W., Razo, J., & Persky, A. M. (2019). Using testing as a learning tool. *American Journal of Pharmaceutical Education, 83*(9), 7324. https://doi.org/10.5688/ajpe7324\

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.