Teaching English Language, Vol. 20, No. 2, Summer & Fall 2026, 231-270.

Teaching English Language Journal
ISSN: 2538-5488 – E-ISSN: 2538-547X – http://tel.journal.org
© 2026 – Published by Teaching English Language and Literature Society of Iran

Please cite this paper as follows:
Ansari, F., Ahmadi Safa, M., (2026). Test fairness in test context framework: Context matters!. *Teaching English Language, 20*(2), 231-270.
https://doi.org/10.22132/tel.2026.550960.1992

Research Paper

Test Fairness in Test Context Framework: Context Matters!

Fatemeh Ansari
Ph.D. Candidate, Department of English Language, Humanities Faculty, Bu-Ali Sina University, Hamedan, Iran
Mohammad Ahmadi Safa[1]
Professor, Department of English Language, Humanities Faculty, Bu-Ali Sina University, Hamedan, Iran

Abstract
High-stakes English standard tests like TOEFL or IELTS play a key role in making some life-changing decisions regarding people's immigration, university admissions, and employment opportunities. Given the profound consequences of these tests, ensuring their fairness is of critical importance. However, despite the multifaceted nature of the test fairness construct, existing literature has predominantly addressed it at micro-level aspects, often overlooking broader social and cultural implications of test administration, and the influence of test contextual factors. Against this backdrop, the current study aimed to develop a context specific test context framework in the Iranian EFL setting. Building upon Kunnan's Test Context Framework (TCF, 2008b), the researchers held interviews with 30 standardized high stakes English general proficiency test takers and TEFL educationalists in this qualitative study. Following template analysis procedure, the content analysis of the interview data resulted in the expansion of TCF to a Revised Test Context Framework (RTCF). The RTCF identified 14 distinctive context types along with their corresponding constituents. The proposed RTCF offers implications at

[1] Corresponding author: m.ahmadisafa@basu.ac.ir

multiple levels for all beneficiaries of high stakes language testing: it provides policy makers with a framework to evaluate test fairness across a variety of contexts, guides researchers in probing into underrepresented contextual variables, such as sociocultural dimensions of language use, and equips test developers as well as specialists with the knowledge to design context-sensitive fair testing practices in high-stakes language testing.
Keywords: Test Fairness, Test Context Framework (TCF), Revised Test Context Framework (RTCF), English Standardized Tests

Received: October 3, 2025
Accepted: February 13, 2026

1. Introduction

While the decisive role of large-scale high stake English proficiency tests in shaping educational and professional future is widely recognized, how the fairness of their application is influenced by the broader contextual factors remains underexplored. This has been mainly due to the fact that language testing researchers have traditionally limited the studies to micro level analyses, overemphasizing the psychometric properties of the tests and have been relatively ignorant of broader sociopolitical, cultural, and economic contexts of the tests (Kunnan, 2008b). Kunnan (2008b) proposes the Test Context Framework (TCF) as a tool to recognize how contextual factors, including the primary test contexts and their corresponding micro-contexts influence test development, administration, use and consequences. Additionally, he adds that depending on the community and the local context as well as the purpose of testing, a range of contexts might emerge and interact in various manners and extent. It is also noteworthy that any fairness framework must be grounded in the realities of its broader implementation contexts to ensure its power and effectiveness (Kunnan, 2010). As Shohamy (2001) demonstrated, these contexts, with the stakeholders involved, can adversely influence test takers, educational institutions, parents, and the community as a whole.

Drawing on Kunnan's (2008b) TCF, the current study tried to respond to two key limitations mentioned in the literature in this regard. First, the TCF fails to establish contextual priorities to validate researches (Karami, 2013). Second, as Kunnan (2010) asserts, local contexts define the characteristics for the manifestations of contextual factors, yet these dimensions remain underexplored. To address these gaps, the present study aimed to develop a Revised Test Context Framework (RTCF) tailored to the context of Iran. The study sought to identify and categorize context- specific factors that influence the fairness of high-stakes English language tests. From a context-sensitive perspective, the present study set out to examine tests and their implementation from a wider contextual perspective to figure out how tests might benefit or harm individuals and communities, and in what ways these influences might be mediated by the local context factors.

2. Literature Review

Traditional language testing has primarily concentrated on technical psychometric aspects of tests, often overlooking prominence of the test contexts and their implications for test takers. Scholars such as Shohamy (2001) have long argued that the context and the contextual factors play significant roles in devising tests and shaping their use and consequences.

In line with this criticism, language test fairness has mainly been narrowly conceptualized through "a nothing-beyond-the-test" approach, which overlooks the context, purposes, and consequences of tests use (Hamid et al., 2019, p.6). To address this gap, Kunnan (2008b, 2009, 2010) called for a more "beyond-the-test approach" which incorporates the contextual aspects into test fairness. He asserts that a de-contextualized testing framework does not capture the complexities of all testing situations.

To operationalize this broader perspective, Kunnan (2008b) devised TCF to elaborate on the contextual aspects of tests and complement his Test

Fairness Framework (TFF). The TCF tapped into five core aspects of political and economic, educational, social and cultural, technological and infrastructural, and legal and ethical dimensions.

Kunnan (2009) applied TCF to examine the U.S. Naturalization Test and contended that the test failed to reach the aims of civic nationalism or social integration. He further maintained that the test enforced burdens on non-English speaking immigrants, imposed challenges on their citizenship process and contributed to racial and ethnic discrimination.

With respect to educational context, Kunnan (2008b) mentioned Stanford Achievement Test (SAT-9) that was administered to grade 2-11 students in California. Highlighting contextual concerns, he contended that the test had negative consequences such as test-directed teaching which led to test-driven curriculum and distorted true teaching and learning process.

Further, technological and infrastructure context was described to be relevant due to rapid expansion and use of technology in testing. This aspect of testing addresses concerns like equal access to technology and knowledge of using it (García-Peñalvo et al., 2021; Kunnan, 2020). Likewise, economic factors make another contextual element that impact test fairness (Kunnan, 2018; Shohamy, 2007; Van der Heijden, 2013). Echoing this issue, Hamid et al. (2019) mentioned numerous economic benefits of IELTS for test owners while imposing financial strain on test takers.

Legal and ethical issues are also deemed significant in their impacts on test fairness (Kunnan, 2008b, 2016, 2018). Language testing also relies on social constructs that are intertwined within value systems which serve social, cultural, and political aims (McNamara, 2001; McNamara & Roever, 2006).

Despite the contextual concerns discussed regarding TCF, Kunnan (2008b) acknowledged the list is not exhaustive and further aspects of the TCF remain to be identified. Furthermore, depending on the purpose and application of a

test, a range of contexts within the TCF may be relevant. In practice, it is likely that one or two contexts may either fully overlap or be only partially involved. Additionally, the influence of these context types largely depends on the motivations and the reasons behind tests' development, administration, scoring, reporting, research, and utilization within a particular community.

Moreover, the primary contexts involved in a specific testing situation may differ from those initially indicated in the TCF. In particular communities and situations, depending on local conditions, the main contexts involved might be social and political, economic and technological, or legal and political, among others (Kunnan, 2008b).

These significant context types should inform the evaluation of motivations, expectations, successes, and failures of tests. Further, as Karami (2013) states, Kunnan in his TCF framework does not establish priorities among the context types. As no two contexts are identical, it is not logical to expect testing contexts to share precisely the same characteristics. Therefore, applying a test without contextual considerations might distort the validity results.

In response to these concerns, and following Kunnan's (2010) call for identification of priorities within particular contexts, it seems necessary to probe into the immediate and most significant contexts of testing influential on test fairness from the view of their prominent stakeholders i.e. teachers and test takers.

In the context of Iran, Foreign Language Education Policy (FLEP) determines language testing practices. Nonetheless, scholars (e.g., Davari & Aghagolzadeh, 2015; Kiany, et al., 2010) have reported inconsistencies among FLEP documents as well as disparities between these documents and English Language Teaching (ELT) practices. Tajeddin and Chamani (2020) further argued that the FLEP developments in Iran are solely determined by

policymakers, excluding other stakeholders such as teachers, students, and school staff.

In higher education context in Iran, Ph.D. candidates are required to pass one of the credible English language proficiency tests of IELTS and TOEFL, or either one of two standard English tests designed and administered by the Iranian Measurement Organization including the Ministry of Science, Research and Technology (MSRT) test and the Test of Language by the Iranian Measurement Organization (TOLIMO). Also a significant body of job applicants are expected to present their English proficiency documents (e.g., TOEFL and IELTS certificate).

Despite the prevalent use and stakes of these tests in Iranian educational and occupational contexts, a review of relevant literature underscored a lack of sufficient studies on context-specific dimensions, based on the scope and sources reviewed by the researchers. In addition, while Kunnan's (2008b) TCF offers a valuable model to investigate such influences, further contextual dimensions remain to be discovered, as acknowledged by Kunnan himself. Building on these insights, the current study set out to examine the contextual aspects most relevant to Iranian English language testing from the perspective of English test takers and educationalists. It also endeavored to provide a more exhaustive understanding of how contextual factors contribute to test fairness. More specifically, this study sought to answer this question: what are the features of an Extended Test Context Framework (ETCF) in Iranian English as a foreign language standard tests context?

3. Method

3.1 Participants

A sample of 30 EFL teachers and non-teacher Ph.D. students either from English or non-English majors who had taken standardized English language tests participated in this study. EFL teachers were included due to their

experience in test design and administration, while Ph.D. students were included due to the mandatory requirement for every Ph.D. student in Iran to present their English proficiency mastery certificate prior to their comprehensive exam.

The participants' age ranged from 26 to 48 (M =37.37, SD=5,63), and the sample included 23 males and 7 females. Announcements for voluntary participation were issued via WhatsApp, Telegram, and also in-person at Shahed University in Tehran. Upon agreement, semi-structured interviews were held with the volunteers. Demographic details are provided in Table 1.

Table 1
Demographic Information of the Participants

Gender			Educational Major			
	Male	Female	TEFL	Engineering	Math	Medicine
N	23	7	6 (3 PhD, 3 MA)	15	4	5
%	76,67	23.33	20 (10 PhD, 10 MA)	50	13.33	16.67

3.2 Instruments and Materials

The study employed a researcher-developed semi-structured interview to explore contextual factors influencing the fairness of English standard tests. Following a comprehensive literature review of fairness frameworks, particularly Kunnan's (2008b), an initial list of 11 tentative contextual elements effective on test fairness were identified. Based on these, 15 open-ended interview questions were posed to elicit the participants' perspectives and lived experiences regarding contextual elements that influence test fairness. To ensure content validity, the questions were reviewed by two TEFL university professors specializing in testing and assessment studies. Their feedback led to the refinement of the items in terms of clarity and relevance.

To enhance response validity and ensure participants' comfort, the interview questions were constructed in the participants' native language i.e.,

Farsi. The questions addressed 11 key factors i.e., social, cultural, economic, ethical, legal, technological, political, educational, religious, gender, and racial, across various stages of test development, administration, scoring, and use. The semi-structured format of interview allowed for flexibility and provision of follow-up questions, enabling the participants to elaborate on the contextual concerns that might not have been included in the original items. This led to the emergence of themes which were absent from the initial codes.

3.3 Data Collection Procedure

Prior to data collection, participation requests were sent to potential candidates via WhatsApp and Telegram, ensuring voluntary participation. To safeguard ethical considerations, the candidates were briefed on the study objectives and assured of data anonymity, confidentiality, and exclusive research use of the collected data.

Upon receiving their consent, interview questions were sent, and responses were collected in Farsi via voice messages. The participants were requested to elaborate on their understanding of the contextual factors influencing test fairness as outlined in the items, with follow-up inquiries to illuminate any ambiguities.

3.4 Data Analysis

The collected data were transcribed and content analyzed. The analysis followed an iterative and reflexive process, integrating both deductive and inductive coding process as the researchers identified and coded recurring themes and patterns in the data (Dörnyei, 2007). Using template analysis which is a form of thematic analysis and hierarchical coding method, the researchers interpreted data. Based on a structured template, they refined codes as new information emerged (Brooks et al., 2015). As Crabtree and Miller (1999) acknowledge, template coding involves establishing codes prior to embarking on data analysis, often based on previous research or preliminary review.

Although template analysis may seem different from other qualitative approaches, Dörnyei (2007) notes that researchers rarely commence data analysis without initial ideas. Additionally, a predefined coding template allows structured interpretation of substantial volume of text in a focused and time efficient manner while creating connections between different data extracts (Dörnyei, 2007).

As mentioned earlier, prior to the interview phase, to the extent possible for the researchers a comprehensive and thorough literature review was conducted that resulted in the development of a template codebook with 11 predefined codes. Next, interview phase followed and after transcription of the interview data, it was meticulously reviewed to identify the recurring themes, which were labeled as initial codes. Then, to identify more abstract commonalities as second-level coding, similar and related initial codes were clustered under broader categories. The coding process was rigorously cross-checked, with ongoing revisions to ensure alignment between the data and the broader categories. While the analysis was guided by the initial themes, it remained open to emergent themes leading to expansion of the template to 14 second-level codes. These additional codes did not result from splitting the existing codes but indicated distinct recurring themes that were not fully represented by the initial coding template.

To ensure coding reliability, a second coder independently reviewed the codes and the equivalent translation of the codes. Inter-coder agreement reached 85 percent, surpassing the recommended 80 percent threshold (Miles & Huberman, 1994). Discrepancies were jointly discussed and refined to ensure all final codes represented shared interpretations and were rooted in the data.

To ensure the trustworthiness (Dornyei, 2007) of the findings, Creswell's (2007) validation strategies were applied. First, the researchers kept a thick

description of the setting, participants and the procedure in mind to maximally assure the transferability of the findings. Moreover, inter-coder agreement was maximally aimed at to ensure the stability and consistency of findings, thereby enhancing the reliability or- its qualitative counterpart i.e. dependability (Dornyei, 2007). It is also noteworthy that this qualitative research was conducted as an initial phase of a broader mixed-methods investigation, which led to the emergence of a subsequent quantitative phase. Through data triangulation (including two rounds of interviews and questionnaires), the researchers attempted to collect converging evidence from multiple sources to ensure the credibility of the findings (Dornyei, 2007). Nonetheless, it is important to note that the present manuscript focuses solely on the qualitative findings.

4. Results

To explore the features of an extended test context framework in Iranian English as a foreign language standard tests context, this section presents the main codes and categories emerged from the qualitative data. The following tables display the initial codes and subcategories extracted from the data concerning different contexts that influence test fairness.

Table 2 presents the initial codes, subcategories and sample interview data concerned with *Social Context* that influence the fairness of English standardized tests.

Table 2

The Initial Codes, Subcategories and Interview Data for Social Context

Initial Codes	Subcategories	Frequency	Samples of Interview Data
1.Social Values	a. Educational Values	13	Positive and negative attitudes of individuals about education might encourage or discourage them to take these tests. Credentialism encourages individuals to pay for the sources and courses.
	b. Career Values	8	Your payment depends on the English degree you have. The type of English degree you have influences your career chances.
	c. Test Values	9	Many people believe English tests are difficult and this fear prevents them from taking these tests.
2. Social Status	-	9	Individuals from higher levels of social class succeed more in these tests.

***Note**:* Frequency refers to the number of participants who mentioned each code.

Table 2 suggests that the interviewees believed in the influence of social values in taking these tests. They also highlighted the significance of social class or status labels of these tests on encouraging individuals to go for the tests. Closely related to the *Social Context* is *Cultural Context*, which is described and displayed in Table 3 below.

Table 3

The Initial Codes and Interview Data for Cultural Context

Initial Codes	Frequency	Samples of Interview Data
1. Culture Free Items	14	Due to cultural differences in societies, test items must be free of cultural points. Items must be designed in a way that all test takers feel culturally respected.
2. Inefficient Cultural Sources	9	The instructional sources in non- English countries are not suitable for teaching English culture.
4. Cultural View	8	Item designers must be aware of their potential cultural biases and guided to control them in the testing process.

As reflected in Table 3, the respondents pointed to the sensitive role of culture in designing items, instructional sources and the cultural views

stakeholders have regarding these tests. Beyond cultural considerations, the participants referred to the influences of *Economic Context* on test fairness (Table 4).

Table 4

The Initial Codes, Subcategories and Interview Data for Economic Context

Initial Codes	Subcategory	Frequency	Samples of Interview Data
1.Costs	a. Sources Costs	8	The high costs of sources prevents test takers from taking standard tests
	b. Courses Costs	10	Due to high costs of courses, many people are not able to participate in courses.
	c. Mock Tests Costs	9	Some might not have the chance to take mock tests due to financial limitation.
	d. Standard Tests Costs	17	High costs of international tests cause abundant stress among test takers influencing performance.
	e. Equipment Costs	10	Due to financial problems, many individuals lack equal access to equipment. The internet cost is high.
2.Government Financial Support	-	13	Due to financial problems in country level, governments do not provide enough free or cheap access for test takers to equipment, sources, courses, mock tests and standard tests.

As shown in Table 4, the interviewees stated that due to the economic challenges, both individuals and governments are not always able to access and provide opportunities for language teaching and testing. In addition, this factor and its related items were one of the most frequently mentioned among all the thematic categories and subcategories.

Turning to ethical aspects of testing, Table 5 illustrates the initial codes and interview data extracted concerning the *Ethical Context.*

Table 5

The Initial Codes, Subcategories and Interview Data for Ethical Context

Initial codes	Subcategory	Frequency	Samples of Interview Data
1. Equality	a. Equal Scoring	10	All test takers should enjoy equal scoring.
			All scorers or interviewers must be instructed about equal scoring
	b. Equal Administration	12	All test takers should enjoy equal administration of tests.
			Equal administration of tests is a precondition for fairness.
			Test takers' performance in different administration conditions cannot be comparable.
	c. Equal Reporting	8	Equal reporting of results raises trust in these tests among test takers.
2. Respecting and Protecting Test Takers	a. Personnel's manner	12	Administrative personnel must treat test takers in a calm and fair manner.
	b. Informed Consent	9	Test takers must be fully informed about the aims, procedures, risk and benefits of the study prior to their participation.
	c. Confidentiality	14	Test providers and authorities must keep test takers' personal data, responses and scores confidential.
	d. Preventing Harm	10	The results of tests should not cause test takers any harm.
3.Transparency	a. Transparent Reporting	9	It is the right of test takers to be clearly informed about their test result.
	b. Transparent Use	7	The ways and goals to use test results must be clearly stated by the authorities.
	c. Transparent Standards	10	Participants must be announced about the process and criteria of testing.

As illustrated in Table 5, the respondents asserted that ethics in testing extends beyond procedural equality. It involves respectful treatment of test takers, protecting their rights as well as transparency throughout the process of testing.

In addition to the discussed contexts, Table 6 displays the initial codes regarding *Legal Context*.

Table 6

The Initial Codes and Interview Data regarding Legal Context

Initial codes	Frequency	Samples of Interview Data
1.Introducing and Enforcing Standards by Setting Up Laws	13	The law should announce and enforce standards concerning the design, administering, scoring of standard tests and reporting their results. There must be some laws concerning the confidentiality of individuals' data and safety of tests.

With regard to *Legal Context*, the respondents emphasized the significance of announcing and enforcing clear standards in testing processes as well as taking measures against violations. According to the participants, such laws can protect test takers' data and right in addition to ensuring procedural consistency and accountability.

Beyond the earlier themes, Table 7 presents the *Technological Context* as another contextual source of impact identified in the data.

Table 7

The Initial Codes and Interview Data for Technological Context

Initial Codes	Frequency	Samples of Interview Data
1.Technological Access	26	All individuals must have access to technology to prepare for and take tests. The low quality of the technological equipment influences test takers' performance and causes stress for them.
2.Technological Literacy	13	All individuals must have enough knowledge on how to use testing equipment.

As shown in Table 7, the respondents stressed the importance of equipping test takers with sufficient technological access, and literacy. The necessity to provide test takers with high quality equipment and stable internet connectivity

was also emphasized as a major challenge. Additionally, this factor emerged as one of the most mentioned barriers to demonstrating true knowledge, highlighting its significance.

Educational context was the following context type extracted from the data indicated in Table 8.

Table 8

The Initial Codes and Interview Data regarding Educational Context

Initial Codes	Frequency	Samples of Interview Data
1.Educational Access	25	All test takers must have access to educational sources, courses and mock tests. Test takers need access to varied teaching methods and sources. The quality of online courses and sources influences test takers' learning and subsequently their performance. Test designers also need enough knowledge and training on developing tests.
2.Government Support	9	In some countries, education is not supported enough by government. Governments need to hold free or cheap preparatory courses.

With respect to this context, which was among the most frequently mentioned ones, the respondents pointed to the significance of providing all individuals with equal access to high-quality education. Moreover, they noted that governmental support for English teaching influences individuals' willingness to participate in such tests and their overall performance.

Further, Table 9 presents the initial codes concerned with *Political Context.*

Table 9

The Initial Codes and Interview Data for Political Context

Initial Codes	Frequency	Samples of Interview Data
1. Educational Policies	11	Main educational policies of government influence designing, administering, scoring, reporting and using tests. Educational policies of governments direct books content.
2.Migrational Policies	13	Migration policies deprive some individuals of taking educational or career chances in other countries.
3.Political Relations	20	Political tensions among countries might deprive the people of some countries of access to international English standard tests. Political tensions among countries might reduce the support of governments of these tests.

As reflected in Table 9, the data indicated that policy might influence different stages and aims of testing which entails implications for individuals' educational and occupational futures. These impacts which are often unrelated to test takers' ability and skill, were perceived to influence tests fairness. For instance, imposing sanctions on their country was cited as a critical political challenge, particularly with Iranians. Such limitations restrict equal access to international education, sources, websites and funds which influence individuals' preparation for and performance as well as evaluation in international tests.

The next context that emerged from the data was named as *Religious Context* (Table 10).

Table 10

The Initial Codes and Interview Data for Religious Context

Initial Codes	Frequency	Samples of Interview Data
1. Religious Neutrality and Respect	10	Items must not address any religious points. The interviewers need to have religious free attitudes in interviewing test takers.

2. Religious Limitations and Attitudes	13	In some religions, both men and women do not have the same rights of education. Due to co-educational English classes, some parents do not let their children take part in these classes. Due to religious attitudes of some families, their children might not have access to some types of sources like music or films.

As evident in Table 10, among the other points indicated in the table, the participants underscored that in designing test items, religious points must be avoided. Furthermore, the test takers must feel respected during the testing process.

Another contextual factor that accordingly needs to be considered was called *Gender-Related Context*, as represented in Table.

Table 11

The Initial Codes and Interview Data for Gender-Related Context

Initial Codes	Frequency	Samples of Interview Data
1. Gender Attitudes	12	In some societies, religions, and culture, women do not have the same rights for education as men. The negative attitudes in societies about one gender influences their educational and occupational prosperity.
2. Gender Free Items Design	10	All items must be gender free. They must not be familiar to a particular gender.

Regarding Table 11, the respondents pointed to gender-related attitudes that influence test takers' access to education and test preparation, potentially impacting their performance. Also they highlighted that cautions about gender influences must be taken into account while designing tests.

Alongside the prior contexts, Table 12 displays the initial codes and interview data concerned with *Racial Context*.

Table 12

The Initial Codes, Subcategories and Interview Data regarding Racial Context

Initial Codes	Subcategory	Frequency	Samples of Interview Data
1. Racism Free Items	-	17	Items must be free of racist topics. Items content must not be familiar to a particular race.
2. Racism Free Scoring	-	9	Interviewers might be influenced positively or negatively by the interviewees' race. Scorers must be aware of their potential racist biases.
3. Racial Access	a. Access to Education	10	It is unfair for races to be denied access to education.
	b. Access to Tests	15	All races and nationalities must have equal access to tests and sources.

Clearly as reflected in Table 12, the respondents underscored the sensitivity of racial factor in different stages of testing. They expressed racial concerns about designing items, scoring and access in addition to unequal access to education and tests across racial groups.

In addition to the earlier themes, Table 13 displays *Geographical Context* as another contextual element.

Table 13

The Initial Codes, Subcategories and Interview Data for Geographical Context

Initial Codes	Subcategory	Frequency	Samples of Interview Data
Geographical access	a. Geographical Access to Tests	14	All cites and countries must have the facilities to hold these tests.
	b. Geographical Access to the Internet	15	All cites and countries must have equal access to the internet.
	c. Geographical Access to Education	20	The public authority needs to provide appropriate education in all areas.

As demonstrated in Table 13, providing equitable access to testing, education and internet across all geographical areas was highlighted as a necessity to fair evaluation of tests.

Alongside the earlier tables, Table 14 displays the data concerned with the *Security Context*.

Table 14

The Initial Codes, Subcategories and Interview Data for Security Context

Initial codes	Subcategory	Frequency	Samples of Interview Data
1.Preventing cheating	a. Test Proctoring	15	Test proctors must be aware of the rules and process of protecting tests.
	b. Preventing Question Leak	8	Question leak decreases the public trust in these tests.
	c. Equipment	17	Different equipment to recognize cheating must be used.
	d. Identity Verification	5	All test sites must be equipped with identity recognition devices and staff.
2. Protecting Data	a. Confidentiality	11	Test designers and givers' identity must be kept confidential.
	b. Anonymous Scoring	7	Test papers and results must be scored anonymously.

Table 14 underscores the significance of ensuring tests security which was one of the most prominent and frequently expressed concerns.

The ultimate context type which emerged from the data was *Administration Context*. Table 15 enlists the codes, subcategories and some sample data related to this context.

Table 15

The Initial Codes, Subcategories and Interview Data for Administration Context

Initial codes	Subcategory	Frequency	Samples of Interview Data
1. Time	-	8	In the case of every disruption in the process of testing, the time should be accounted.
2.Physical Setting	a. Ventilation Systems	8	Test takers might not be able to concentrate due to inappropriate temperature.
	b. Seats	5	Chairs must be comfortable enough to sit on for some hours.
	c. Lighting System	5	Improper light might cause fatigue in test takers.
	d. Noise	7	Test takers concentrate better in quiet places.

Regarding *Administration Context* in Table 15, the respondents raised several concerns regarding physical setting. In this respect, respondent 8 stated "I still remember how cold I felt during my TOLIMO. I was so cold that I was not able to think well". Also respondent 2 noted "I was distracted by the noise test proctors made while walking".

Additionally, concerning time, respondent 9 explained "since there was around a 30-minute delay by the authorities to begin the testing process, I felt tired and could not concentrate my mind on the listening part". These statements highlighted the importance of administration context in facilitating the condition to support test takers' concentration on tests.

Overall, as are presented above, 14 second level codes concerning the contextual factors influencing the fairness of English standard tests were extracted from the data. Table 16 illustrates these second level codes.

Table 16

The Second Level Codes for Contextual Factors

Second Level Codes													
Social Context	Cultural Context	Economic Context	Ethical Context	Legal Context	Technological Context	Educational Context	Political Context	Religious Context	Gender –related Context	Racial Context	Geographical Context	Security Context	Administration Context

As evident in Table 16, it should be noted that the contexts of *Ethical*, *Religious*, *Gender- Related*, *Racial*, *Security* and *Administration* were either overlooked or insufficiently represented in Kunnan's (2008b) original framework though they appeared frequently in the current study context. Such additional context types led to a redefinition of test context framework named as Revisited Test Context Framework (RTCF).

To indicate which contexts were the most salient in the data, the frequency and percentage of all 14 contexts are presented in descending order in Table 17.

Table 17

Frequency and Percentage of RTCF Contextual Categories

Contextual Category	Frequency	Percentage (%)
Ethical	101	16.61
Economic	67	11.02
Security	63	10.36
Racial	51	8.39
Geographic	49	8.06
Political	43	7.07
Social	39	6.41
Technological	39	6.41
Educational	34	5.59

Administration	33	5.43
Cultural	31	5.10
Religious	23	3.78
Gender-related	22	3.62
legal	13	2.13%

As the frequency count and the percentages indicate, *Ethical*, *Economy* and *Security* contexts emerged as the most frequently mentioned context types, highlighting the participants' sensitivity to these aspects in testing contexts. In contrast, *Legal*, *Gender-Related* and *Religious* contexts were noted less often, while still contributing to the RTCF.

5. Discussion

Building upon Kunnan's (2008b) call for a more elaborate test context framework, this study sought to expand the four-factor TCF model by identifying contextual factors that affect the fairness of standardized English proficiency tests from EFL teachers and Ph.D. students' perspectives. The findings led to the nomination of 14 major factors, each of which is discussed in the following section.

With respect to the *Social Context,* the participants highlighted the existing values in societies regarding *Educational*, *Career* and *Tests Values*. As noted by Messick (1989) language tests inherently embody values, and it is necessary to appreciate social values in order to interpret test scores. The participants emphasized the significance of society and family attitudes in shaping *Educational Values*, such as credentialism which compels individuals to invest substantial time, effort, and money in test preparation. Regarding *Career Values*, the participants highlighted the role of educational qualifications in enhancing occupational prospects and salary. Further, with respect to *Test Values*, the recognized status of standardized tests fosters trust in them ultimately normalizing these tests and restricting public understanding of their underlying social structures (Garrison, 2020). The participants' reflections on

Social Context seem to be in alignment with the social context of Kunnan's (2008b) TCF, in particular his emphasis on washback as a social consequence.

The findings also underscore the influence of *Social Status* of individuals on testing, implying that individuals from higher socioeconomic status are more likely to participate and succeed in these tests. Nevertheless, Shohamy (2010) reported a paradox wherein some marginalized groups often value standard tests and oppose their abolishment. Despite this, these tests favor individuals who already hold advantages, thereby contributing to existing social hierarchies (Garrison, 2020). This code can be related to Kunnan's (2008b) component of absence of bias in the TTF and TCF, in which different group membership may result in varied performance, contributing to test bias. Overall, as highlighted by Khan et al. (2025), these themes underscore how social expectations shape priorities and experience of fairness

Concerning *Cultural context*, the respondents referred to *Culture-Free Items*, *Inefficient Cultural Sources*, and *Cultural Views*. These concerns align with Pennycook (2017) who argues that English language testing is historically rooted in colonial English disregarding local languages. Similarly, Pusawati (2014) named TOEFL, the content of which requires familiarity with U.S. culture, as a test which instantiates Western cultural dominance. The participants' perspectives on the concept *Culture-Free Items* is in line with Kunnan's (2008b) cultural context in TCF, in which he emphasized the absence of content bias to test takers from particular backgrounds which can be related to culture too. Nonetheless, he did not identify the components of *Cultural context* and elaborate them.

The respondents also highlighted the inadequacy of instructional materials to address cultural dimensions. In this regard, Pennycook (2017) contends that the worldwide spread of English through standardized tests emphasizes its cultural dominance, positioning Western norms as standards and disregarding

local languages. This oversight is evident in EFL textbooks, which represent diverse cultural values sporadically (Cook, 2008), which can pose challenges for non-native speakers to succeed (Pennycook, 2017). Taken together, referring to Zhaleh et al. (2025) test takers' experiences are filtered through sociocultural background, highlighting the need for context- sensitive fairness frameworks.

Beyond the contexts discussed, as Shohamy (2010) contends, language testing is embedded in *Educational Context*. Kunnan (2008b) in his TFF mentioned the concept of educational access, elaborating it as test takers' opportunity to learn relevant content and get acquainted with required tasks and cognitive demands. However, the findings of the current study emphasized the importance of access to appropriate educational sources and courses, teachers, varied teaching methods, mock tests and information regarding the type of these tests, categorized as *Educational Access*. In this respect, Shohamy and Pennycook (2019) stressed the significance of equipping test designers with enough knowledge and training on test development. *Government Support* was additionally identified as a crucial factor, as the participants pointed to the influence of state policies on education and testing, which is in alignment with Shohamy (2010). In general, Kunnan (2008b) in his TCF defined the educational, social and cultural context under one general category without identifying its constituents, while in the current study, the findings led to three separate *Educational*, *Social* and *Cultural* contexts, with their corresponding components.

The findings also highlighted the impact of *Economic Context* of language testing, including *Costs of sources*, *Courses*, *Mock Tests*, *Standard Tests* and *Equipment* as well as lack of *Government Financial Support*. In addition, the findings indicated that large scale language testing has turned into a highly profit-making industry. In line with this, Chik and Besser (2011) and Hamid et

al. (2019) noted that the IELTS test-retake policy is mainly profit-driven, a concern expressed by the participants in this study too.

Additionally, Shohamy (2007) maintained that language tests add to economic inequalities by advantaging those who gain access to proper education, often available to the economically privileged individuals. Furthermore, the employment of large-scale standardized testing as a mechanism for resource allocation and accountability of public schools underscores the economic significance of such tests, which Kunnan (2008b) criticizes.

Similarly, Berliner and Biddle (1995) as well as Rothstein (2004) contended that individuals with lower earning power frequently do not enjoy essential educational preconditions, including appropriate healthcare and nutrition. In short, these economic barriers highlight the necessity to approach financial inequalities in test accessibility and preparation.

With respect to *Political Context*, the findings revealed three key concepts as *Educational Policies*, *Migration Policies*, and *Political Relations*. Concerned with *Educational Policies*, they highlighted how governments' educational policies influence testing process through providing access, support, and preparatory resources. As Garrison (2020) and Shohamy (2001) contend standardized tests have allowed public and private sectors to exert their influence on educational priorities and practices by controlling what they must know and learn.

Language tests are used as a mechanism for controlling immigration and refugees' access to social participation (Eades et al.2003; Eades, 2005). Reaffirming this, the findings of this study noted how migration policies shape different stages of testing, influencing individuals' professional and educational future.

Additionally, the participants raised the points that *Political Relations* between countries can either pose challenges or offer test-takers with more opportunities to take tests, impacting educational content, accessibility, and government support. For instance, due to international sanctions imposed on Iran, Iranians' access to international education, resources, and funding has been constrained which is in alignment with Shohamy (2010). Overall, the political dimensions addressed in this study highlight the necessity of assessing language tests through the lens of a TCF in addition to a TFF as underscored by Kunnan (2008b). However, while Kunnan (2008b) treats both political and economic contexts as a unified category, in this study these two contexts are separated into distinct domains, each with its own constituents.

According to Davies (2008), to have fair and accurate tests, in addition to technical soundness, ethical principles must be watched. Also ethics have been considered as the prerequisite for test validity and fairness (Bachman & Purpura, 2008; Kunnan, 2018) as it helps eliminate biases in test design, administration, and scoring, (Davies, 2008). Kunnan (2020) proposed an ethics-centered approach that ensures all individuals must be provided with equal chances to show their abilities and treated with equal respect, a point that is confirmed in the current study as well.

The findings also stressed *Respecting and Protecting Test Takers* in terms of *Personnel Manner*, *Informed Consent*, *Confidentiality* and *Preventing Harm*. This aligns with the ILTA Codes of Ethics and supported by Fulcher and Davidson (2007) as well as Salaberry et al. (2023) who stress the significance of respecting test takers' dignity and humanity and also minimizing harm which can be related to consequential validity. Moreover, ILTA guidelines for practice (2007) considers informed consent, confidentiality and voluntary engagement in testing, as test takers' fundamental rights.

The participants stressed the need for *Transparency* in test *Reporting*, *Use,* and *Standard*. Referring to ILTA guidelines for practice (2007) and Davies (2008) testing institutions must explicitly articulate the test's purpose, the constructs aimed to be measured, methods, as well as scoring and reporting mechanisms. In general, in terms of *Ethical Context*, Kunnan (2008b) pointed to taking remedial measures in case of test violations, however, in the current study, some preconditions as equal scoring, administration and scoring, respecting and protecting participants from harm and also transparency were mentioned as the criteria to prevent unethical conduction and use of tests.

Closely related to ethical context, Kunnan (2008b) introduced *Legal Context*. Regarding this context, the participants acknowledged the significance of having clear standards, yet emphasized the role of legal requirements to establish and enforce them. Concerning this, Bachman and Palmer (2010) highlighted the necessity to pass laws to address discrimination challenges, along with procedural issues including appropriate advance notice and instruction concerning necessary knowledge and skills about tests. Likewise, the ILTA (2007) guidelines for practice emphasize the responsibility of test developers and administrators to enact laws for denouncing test misuse, legal actions, and even abolishing a test in the case of its misuse. In this regard, Kunnan (2008b), while taking both the legal and ethical contexts as one concept, discussed three challenges in terms of discrimination, the due process and accommodations required for test takers having disabilities. He took these measures as remedies to tackle the challenges in case of their occurrence, while in the current study, the measures were introduced in terms of *Introducing and Enforcing Standards by Setting Up Laws* as preventive actions.

Concerning *Technological Context*, the technological and educational aspect of language testing, including *Access* and *Literacy* were identified. Regarding *Access*, the participants called attention to the presence of advanced

equipment, stable internet connections, and technological support. They also underscored the significance of equipping test takers with essential technology knowledge prior to testing, coded as *Technological Literacy*. Similarly, Ahmadi Safa and Anasari (2026) and also Kunnan (2008b) highlighted the significance of providing test takers with both access to and awareness of the required technology. Overall, if technology hinders rather than facilitates testing, scores may become confounded with technological access and knowledge deficits which doubt the test validity.

In relation to *Religious Context*, the findings underscore that test items and interview processes should be devoid of religious bias, aligning with Kunnan's (2008b) TFF. Moreover, as noted by the participants, religious beliefs, practices and constraints may impact access to resources, participation in co-educational settings, and attitudes toward interviews and testing as well as personal comfort and perceptions during interviews which were coded as *Religious Limitations and Attitudes*. For instance, concerns related to attire, especially among Muslim women, may pose negative perceptions in the interview process which was also pointed out by Wollack and Case (2016). In this regard, Kunnan (2008b) briefly acknowledged the potential existence of religious context in TCF, however, he did not elaborate it extensively.

Regarding *Gender-Related Context*, the participants highlighted social, cultural, and religious attitudes concerning gender abilities and rights, which might impact test performance besides future educational and occupational opportunities. In this respect, Kunnan (2000) emphasizes that raters' perceptions and biases about gender can influence scoring. For instance, female test-takers may be expected to have a better performance in language tasks due to presumed verbal abilities. Therefore, to ensure fair testing, it is crucial to raise raters' awareness towards such biased orientations.

Moreover, the findings highlight the importance of developing gender-neutral test items. Kunnan (2000) and Shohamy (2001) claimed that gender difference in test performance can stem from biases embedded in test formats and content, such as topic selection, examples, language use, and cultural references that favor one gender. Empirical studies, including Akhavan and Sadeqi (2020), Reardon et al. (2018) and Taylor and Lee (2012) have indicated the effects of test formats on performance of both genders.

Further, the findings highlighted that test items, content and scoring process must be free of racial biases, coded as *Racial Context*. According to Shohamy and Pennycook (2019), test-takers belonging to minority racial and ethnic groups often experience unfamiliar cultural references and linguistic norms in tests. Additionally, informed by the results of such biased language tests, individuals from marginalized racial and ethnic groups may be denied equal access to education and professions, intensifying systemic discrimination (Kunnan, 2008b). In line with this, the participants noted that all individuals should have access to education and testing, regardless of race, a concern categorized as *Racial Access*.

Historical cases such as Australian dictation test (Kunnan, 2008b) and the Army Alpha/Beta tests (Murdoch, 2007) demonstrate the use of language for racial exclusion. Criticizing the Army Alpha/Beta tests, Bond (1924) argued that they described intelligence as a racial construct rooted in Nordic American cultural assumptions. In general, as Shohamy and Pennycook (2019) contend, language tests are not neutral tools but reflect the cultural and racial biases of their designers. While this study distinguishes the constructs of gender and race as distinct concepts, Kunnan (2008b) briefly acknowledges their potential existence without specifying them in his model.

As indicated in the findings, *Geographical Context* also has a key role in test performance, extending beyond mere physical distance to testing sites.

While Kunnan (2008b) defined geographical access as a TFF component, the participants of this study highlighted it as a broader contextual factor since geographical situation is one of the contexts in which test functions.

Concerning this context, Kunnan (2008b) noted the importance of test site accessibility in terms of distance in his TFF, however our participants emphasized that all individuals in geographic regions need to have access to education, and internet as well. Similarly, Sanday (1999) argued geographical isolation negatively impacts test performance due to limited exposure to cultural knowledge, which suggests that cultural bias in tests can be traced back to geographical disparities.

Security Context is found to be another essential component of fair testing context in this study. It incorporated measures of *Preventing Cheating* and *Protecting Data*. The findings identified key cheating concerns, including staff proctoring, preventing question leaks, using appropriate equipment, and verifying test takers' identities.

According to Kunnan (2008b), test security as a component of administration in his TFF refers to breach of security of test materials or test administration including fraud, misrepresentation, cheating, and plagiarism. As reported by Standards for Educational and Psychological Testing (AERA et al., 2014) and confirmed by the findings in the current study, test takers are accountable for their behavior during assessments, such as not indulging in any cheating, fraud, deceit, or plagiarism. In addition, proctoring was found as an important security measure in this study, consistent with some previous studies (e.g., Kunnan, 2008b; Roe et al., 2023; Wollack & Case, 2016).

Moreover, according to the findings, reliable test administration necessitates rigorous security *Equipment* to prevent and deter answer sharing, item theft, cheating, unauthorized access to test materials, and fraudulent practices. To safeguard test security, *Identity Verification* was also noted

effective, particularly in large-scale tests where technology could be exploited for fraud detection. Further, as noted by Geranpayeh (2014) and reflected in the participants' responses, psychometric examinations, which analyze test performance patterns to identify irregularities, can act as an effective cheating detection strategy.

Additionally, Kunnan (2018) argues that test security extends beyond test administration, involving test users too. Similarly, Shohamy (2001) stressed that practitioners and researchers must pay due attention to both their own safety and that of individuals whom they are professionally responsible for. Consistently, the participants highlighted the significance of *Protecting Data*, including securing personal information of test takers, administrators, users, and designers. Moreover, maintaining confidentiality of responses and test results can prevent manipulation, as raised by the participants. However, as highlighted in the literature, *Protecting Data* is addressed within both ethical and security frameworks (e.g., AERA et al., 2014; Fulcher & Davidson, 2007). Accordingly, in the present study, this concept is put under both categories to reflect its dual significance in ensuring secure and ethical data management.

Further, the findings stressed the impact of *Administration Context* on test fairness, emphasizing the significance of *Time* and *Physical Setting*. As noted in the literature and reaffirmed by the participants, test performance is influenced by various administration factors, and if these factors are not truly controlled, test scores may be reflective of these disparities rather than actual abilities which may distort the validity of the tests (Haladyna & Downing, 2004; McCallin, 2006; Wollack & Case, 2016).

Moreover, according to the findings of this study, and as emphasized by Beheshti and Ahmadi Safa (2023) as well as Haladyna and Downing (2004), timing and time budgeting are essential for fair test administration. This

includes sufficient and equal time allocation, exact start and end times, and considering test takers' peak performance periods.

Regarding the *Physical Setting*, the study data referred to *Ventilation Systems*, *Seat*, *Lighting System* and *Noise* codes. Consistent with earlier investigations (e.g., Beheshti & Ahmadi Safa, 2023; Wollack & Case, 2016), the findings underscored comfortable testing environment including well-lit rooms with suitable temperature.

As noted earlier, in Kunnan's (2008b) TFF, test security was considered as a component of administration. However, in this study the analyses led to separation of these components within RTCF, as they represent the context within which the fairness is enacted and practiced. Additionally, as stated by Kunnan (2018), test security extends beyond test administration, involving test users too. Similarly, protecting data and results from manipulations is considered as an element of test security in the present study. Weir (2005) points out that test developers and users should consider test immediate context to ensure fairness in administration, which suggests test administration as a part of RTCF. Therefore, given Kunnan's (2008a, p. 254) elaboration on test context framework that emphasizes on "why and how a test is commissioned, developed, administered, scored, reported, researched, and used by the community in which the test operates", it can be concluded that test administration can be a distinct aspect of a revised TCF.

The distribution of each context within the RTCF reveals the multifaceted nature of testing and salience of some contexts as experienced by test takers and educators. The predominance of *Ethical* and *Security* contexts highlights the point that the participants are pretty aware of their fundamental right as test takers and procedural integrity. Moreover, their attention to *Economic* context reveals their lived experience of affordability of standardized tests. Furthermore, the significance of Ethical context in the findings is in alignment

with Kunnan's (2020) ethic- centered approach to testing. While contexts such as *Legal*, *Gender-Related* and *Religious* appeared less frequently, their presence highlights the significance of their inclusivity in RTCF. Therefore, the frequency data may aid to confirm the comprehensiveness of the RTCF as well as offering a heuristic tool for planning and prioritizing intervention practices.

Overall, to summarize the revisions made into TCF, it is important to highlight that although Kunnan (2008b) in his TCF acknowledged the importance of context and laid the foundation for recognizing the influence of contextual factors on test fairness, he provided a broad and brief discussion of contextual factors, combining them into four general categories rather than discussing them individually without specifying the rational for combination. In contrast, the current study tried to offer a more empirically grounded picture of a TCF by differentiating contextual factors into 14 distinct contexts, each with its own constituents. Further, Kunnan did not unpack the complexity of each contextual factor, however, this study tried to advance a more nuanced understanding of each distinct context and better capture the multifaceted reality of context. This study attempted to respond Kunnan's (2008b) acknowledgment about the incomplete nature of his TCF and also operationalize his call for context driven evaluation of tests by identifying the sub-components that shape test takers' performance. The RTCF contains the 14 second-level codes of *Social*, *Cultural*, *Economic*, *Ethical*, *Legal*, *Political*, *Technological*, *Educational*, *Political*, *Religious*, *Gender- Related*, *Racial*, *Security* and *Administration* contexts. In addition, the contexts of *Ethical*, *Religious*, *Gender- Related*, *Racial*, *Security* and *Administration* that were either absent or underdeveloped in Kunnan's (2008b) original framework are added to the RTCF.

6. Conclusion and Implications

Building upon Kunnan's (2008b) TCF, this study resulted in the extension of his broad theoretical foundations through proposing a RTCF to link the TFF to the context of actual language use. While the TCF offers a foundational framework for evaluating language tests, it treats its constructs as a group of unified categories. In contrast, the RTCF presents a broader and empirically grounded model, identifying 14 distinct contexts with their corresponding constituents. Further, our findings confirm earlier studies suggesting that tests, their status in societies and test takers' educational and testing experiences might exert influence on their performance and ,in turn, affect test fairness (Kunnan, 2020, 2016, 2009; Shohamy & Pennycook, 2019; Weir, 2005).

Expanding Kunnan's (2008b) perspective, the RTCF offers a lens through which stakeholders can design and evaluate fairer tests informed by contextual dimensions. Educational testing researchers can benefit from this framework to analyze the interplay between contextual factors (e.g. geographical access, racial context and political orientations) and test design and use. Also, test developers and practitioners are urged to address contextual factors, including test administration, security and technological access, when designing or adapting tests. Teacher professional development programs can employ the RTCF to train educators how to identify and mitigate contextual factors that interfere with test performance. Furthermore, policy makers are advised to utilize the RTCF to ensure that testing policies prevailing in governments do not contribute to unfairness of high-stakes tests. Moreover, using this framework, educators will be equipped with knowledge and tools to help test takers deal with contextual challenges, such as unfamiliar cultural content. Most importantly, the RTCF helps all major stakeholders recognize test takers not as sole recipients of tests, but as those whose life might be shaped by the broader contexts in which tests are embedded. As Shohamy (2001) contends,

these contexts, along with the stakeholders involved, can yield crucial influences on test takers, educational institutions, parents, and the broader community. Overall, the RTCF attempts to reframe fairness from an abstract notion into a context-grounded practice.

Finally, as a word of caution, due to the small sample size of this study, the finding need to be cautiously applied and re-examined in other contexts. In addition, the framework was developed using the perspectives of teachers and PhD students or holders, whose insights -though valuable-may not have fully represented the intricacies of the RTCF. Therefore, future studies should include the voices of English standard test developers to capture a more comprehensive framework. It seems logical for future investigations to include a more representative and inclusive sample of stakeholders and apply triangulated methods of data collection to present a more vivid and comprehensive picture of all potential context types and their constitutional sub-factors. Furthermore, the gender imbalance in the sample (23 males, 7 females), which was due to their unwillingness to participate in the study, may have influenced the findings concerned with gender-related context. Hence, further investigations with more balanced representation are required to capture the complexity of the concept.

References

Ahmadi Safa, M. Ansari, F. (2026). Test Fairness in Online Assessment: Insights from Iranian EFL Teachers Perspective. *Teaching English as a Second Language Quarterly, 45*(1), 81-108. https://doi.org/10.22099/tesl.2025.52797.3395

Akhavan Masoumi, G., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: The role of gender. *Language Testing in Asia, 10*(1), 2. https://doi.org/10.1186/s40468-020-00099-x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014).

Standards for educational and psychological testing. American Educational Research Association.

Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). John Benjamins.

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

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or door openers? In B. Spolsky & F. M. Hult (Eds.), *Handbook of Educational Linguistics* (pp. 456–468). Blackwell Publishers. https://doi.org/10.1002/9780470694138.ch32

Beheshti, Sh., & Ahmadi Safa, M. (2023). Reconceptualization of Test Fairness Model: A Grounded Theory Approach. *Iranian Journal of Language Teaching Research, 11*(2),119–146. https://doi.org/10.30466/ijltr.2023.121333

Berliner, D., and Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Perseus Books.

Bond, H. (1924). Intelligence tests and propaganda. *Crisis, 28*(2), 61–64.

Brigham, C. C. (1975). *A study of American intelligence*. Kraus Reprint Co.

Brooks, J., McCluskey, S., Turley, E., & King, N. (2015). The utility of template analysis in qualitative psychology research. *Qualitative Research in Psychology*, *12*(2), 202–222. https://doi.org/10.1080/14780887.2014.955224

Chik, A., & Besser, S. (2011). International language test taking among young learners: A Hong Kong case study. *Language Assessment Quarterly, 8*(1), 73–91. https://doi.org/10.1080/15434303.2010.537417

Cook, V. (2008). *Second language learning and language teaching*. Hodder Education.

Crabtree, B. F., & Miller, W. L. (1999). *Doing Qualitative Research* (2nd ed.). Sage.

Davari, H., & Aghagolzadeh, F. (2015). To teach or not to teach? Still an open question for the Iranian education system. In C. Kennedy (Ed.), *English language teaching in the Islamic Republic of Iran: Innovations, trends and challenges* (pp. 10-19). British Council.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (2nd ed., Vol. 7, pp. 429-443). Springer. https://doi.org/10.1007/978-0-387-30424-3_1911.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies*. Oxford University Press.

Eades, D. (2005). Applied Linguistics and Language Analysis in Asylum Seeker Cases. *Applied Linguistics, 26*(4), 503–526. https://doi.org/10.1093/applin/ami021

Eades, D., Fraser, H., Siegel, J., McNamara, T., & Baker, B. (2003). Linguistic identification in the determination of nationality: A preliminary report. *Language Policy*, *2*(2), 179–199. https://doi.org/10.1023/A:1024640612273

Fulcher, N.G. & Davidson, Fred. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

García-Peñalvo, F. J., Corell, A., Abella-García, V., & Grande-de-Prado, M. (2021). Recommendations for mandatory online assessment in higher education during the COVID-19 pandemic. In D. Burgos, A. Tlili, & A. Tabacco (Eds.), *Radical Solutions for Education in a Crisis Context* (pp. 85–98). Springer. https://doi.org/10.1007/978-981-15-7869-4_6

Garrison, M. J. (2020). Standardized testing, innovation, and social reproduction. In *Encyclopedia of Educational Innovation* (pp.1-7). Springer. https://doi.org/10.1007/978-981-13-2262-4_118-21

Geranpayeh, A. (2014). Detecting plagiarism and cheating. In A.J. Kunnan (Ed.), *The companion to language assessment* (pp. 980–993). Wiley.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17–27. https://doi.org/10.111/j.1745-3992.2004.tb00149.x

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language Testing in Asia. 9*(1), 1-20. https://doi.org/10.1186/s40468-019-0092-9

House, E. R. (1990) 'Ethics of evaluation studies.' In H. J. Walberg & G. C. Haertel (Eds.), *The International Encyclopedia of Educational Evaluation*. (PP.91–94). Pergamon Press

International Language Testing Association. (2007). *ILTA guidelines for practice in English.* ILTA

Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation: An International Journal on Theory and Practice, 19*(2–3), 158–169. https://doi.org/10.1080/13803611.2013.767621

Khan, A., Hassan, N., & Cheng, L. (2025). Investigating the contextual factors mediating washback effects of a learning-oriented English language assessment in Malaysia. *Language Testing in Asia, 15*(20). https://doi.org/10.1186/s40468-025-00359-8

Kiany, G. R., Mirhosseini, S.A., & Navidinia, H. (2010). Foreign language education policies in Iran: Pivotal macro considerations. *Journal of English Language Teaching and Learning*, *2*(2), 49-70.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan, (Ed.), *Fairness and validation in language assessment* (pp. 1-13). Cambridge University Press.

Kunnan, A. J. (2008a). Large scale language assessments. In: Hornberger, N.H. (Eds), *Encyclopedia of Language and Education.* Springer. https://doi.org/10.1007/978-0-387-30424-3_173

Kunnan, A. J. (2008b). Towards a model of test evaluation: Using the Test Fairness and Wider Context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229–251). Cambridge University Press.

Kunnan, A. J. (2009). Testing for citizenship: The U.S. Naturalization Test. *Language Assessment Quarterly*, *6*, 89–97.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*(2), 183–189. https://doi.org/10.1177/0265532209349468

Kunnan, A. J. (2016). Large-scale language assessment. In R. M. Paige & D. L. Lange (Eds.), *Handbook of research in second language teaching and learning* (pp. 34-48). Routledge. https://doi.org/10.4324/9781315716893-34

Kunnan, A. J. (2018). *Evaluating language assessment*. Routledge.

Kunnan, A. J. (2020). A case for an ethics-based approach to evaluate language assessments. In G. J. Ockey & B. A. Green (Eds.), *Another generation of fundamental considerations in language assessmen*t: *A Festschrift in honor of Lyle F. Bachman* (pp. 77-93). Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-15-8952-2_6

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625–652). Lawrence Erlbaum Associates

McNamara, T. (2001). *Language Testing*. Oxford University Press.

McNamara, T. (2007). Language Testing: A Question of Context. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language Testing Reconsidered* (pp. 131–138). University of Ottawa Press. https://doi.org/10.2307/j.ctt1ckpccf.13

McNamara, T. and C. Roever (2006). *Language Testing: The Social Dimension*. Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Murdoch, S. (2007). *IQ: A smart history of a failed idea*. John Wiley & Sons.

Pennycook, A. (2017). *The cultural politics of English as an international language* (1st ed.). Routledge. https://doi.org/10.4324/9781315225593

Pusawati, I. (2014). Fairness issues in a standardized English test for nonnative speakers of English. *TESOL Journal, 5* (3), 555-572. https://doi.org/10.1002/tesj.157

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher, 47*(5), 284–294. https://doi.org/10.3102/0013189X18762105

Roe, J., Perkins, M.,Chonua, G. K., & Bhatia, A. (2023). Student perceptions of peer cheating behaviour during COVID-19 induced online teaching and assessment. *Higher Education Research & Development*, *43(4)*, 1-15. https://doi.org/10.1080/07294360.2023.2258820

Rothstein, R. (2004). *Class and schools: Using social, economic and educational reform to close the Black-White achievement gap.* Economic Policy Institute

Salaberry, M. R., Weideman, A., & Hsu, W.-L. (2023).*Ethics and context in second language testing: Rethinking validity in theory and practice.* Routledge. https://doi.org/10.4324/9781003384922

Sanday, P. (1999). On the causes of IQ differences between groups and implications for social policy. In A. Montagu (Ed.), *Race and IQ* (expanded ed., pp. 276–307). Oxford University Press.

Shohamy, E. (2001). *The power of tests: Critical language testing*. Routledge. https://doi.org/10.4324/9781315837970

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education: Principles, Policy & Practice, 14*(1), 117–130. https://doi.org/10.1080/09695940701272948

Shohamy, E. (2010). *The Power of Tests: A Critical Perspective on the Uses of Language Tests.* Routledge.

Shohamy, E., & Pennycook, A. (2019). Extending Fairness and Justice in Language Tests. In C. Roever, & G. Wigglesworth (Eds.), *Social Perspectives on Language Testing: Papers in Honour of Tim McNamara* (pp.29–45). Peter Lang AG.

Tajeddin, Z., & Chamani, F. (2020). Foreign language education policy (FLEP) in Iran: Unpacking state mandates in major national policy documents. *Journal of Teaching Language Skills (JTLS)*, *39*(3.1), 185-215. https://doi.org/10.22099/jtls.2021.38870.2904

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25* (3), 246–280. https://doi.org/10.1080/08957347.2012.68765

Van der Heijden, J. (2013). *Testing skilled migrants' English: Ridiculous and insulting*. Independent Australia. https://independentaustralia.net/australia/australia-display/testing-skilled-migrants-english-ridiculous-and-insulting,5989

Weir, C. J. (2005*). Language testing and validation: an evidence-based approach*. MacMillan Palgrave.

Wollack, J. A., & Case, S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 33–53). Routledge.

Zhaleh,K. , Estaji,M. and Chory,R. M. (2025). Justice and Fairness are not the Same Construct: Evidence from Revalidating the Teacher Classroom Justice Scale on University EFL Students in Iran. *Teaching English Language, 19*(1), 41-80. https://doi.org/ 10.22132/tel.2025.473990.1675

2026 by the authors. Licensee Journal of Teaching English Language (TEL). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution–Non Commercial 4.0 International (CC BY-NC 4.0 license). (http://creativecommons.org/licenses/by-nc/4.0).