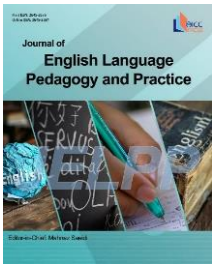

Research Article

Volume 18, Issue 2, (2025) Article 14

Journal of
English Language Pedagogy and Practice
(JELPP)

https://doi.org/10.57647/jelpp.2025.1802.14

Developing a Critical Checklist for High-stakes English Proficiency Tests in Iran

Farzaneh Nouri, Shaban Najafi Karimi*

English Department, QaS. C., Islamic Azad University, Qaemshahr, Iran

***Corresponding author:** sn.najafi@iau.ac.ir

Original Research

Received:
11 May 2025
Accepted:
16 November 2025
Published in Issue:
31 December 2025

Abstract

The present study aimed to develop a critical checklist for evaluating high-stakes English proficiency tests based on critical language testing concept extensively utilized in Iranian higher education context, namely EPT exam, which plays a significant role in the field of language testing. To this end and by adopting mixed-methods design, 66 items were developed based on extensive review of the related literatures and consultation with experts and interviews. The tentative checklist was distributed to 350 participants. The collected data were analyzed by running exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and Cronbach alpha coefficient. The results confirmed the reliability and validity of the newly developed instrument. EFA presented six components and CFA confirmed the statistical support for the six components while all of them had high-reliability estimates. The computed model-fit estimates confirmed the CFA model. In a novel way, this checklist evaluates tests critically to address injustice against minorities that others do not hear their voices due to power imbalances among test parties. It aims to reduce inequalities, remove biased test items, and promote fairness among stakeholders. In addition, this checklist helps test developers make high stakes tests as fair and democratic as possible while not missing any important issues in the process of developing exams from the critical point of view.

Keywords: Checklist; Critical language testing; Democratic assessment; Fairness; Power distribution

Cite this article: Nouri, F., Najafi Karimi, Sh., Developing a Critical Checklist for High-stakes English Proficiency Tests in Iran, *J. Engl. Lang. Pedagog. Pract.*, 18(2) Article 14 (2025). doi.org/10.57647/jelpp.2025.1802.14

1. Introduction

Tests are powerful tools in determining the future lives of individuals in English as a foreign language (EFL) contexts. Some tests are used for making important decisions (Bachman, 2013), and passing or failing them has noticeable consequences. High-stakes tests are clear manifestations of this impact as they can act as "both door-openers and gatekeepers" (Bachman & Purpura, 2008, p. 456) and influence many aspects of test takers' lives. In Iran, language tests are often used for graduation and granting scholarships (Keyvanfar & Dadfarma, 2013) and are mostly administered in traditional pen-and-paper formats such as multiple-choice, true–false, matching, and short-answer items. Although multiple-choice tests are popular because they

are easy to administer and objectively scored, Shohamy (2001) argues that traditional tests are detached from individuals, society, test purposes, and outcomes, and fail to align with test goals and test takers' emotions. Messick (1989) similarly emphasized the need to move from traditional testing toward more ethical practices that attend to test impact, social consequences, and uses of tests.

With the advent of critical language testing (CLT), the "rights of test-takers" (Shohamy, 2001a, p. 115) came to the foreground, challenging the earlier view of test takers as passive recipients of decisions imposed by powerful organizations. High-stakes tests empower policymakers to enact and verify educational policy while the public often accepts and trusts these tests uncritically, enabling dominant groups to perpetuate their control (Bourdieu, 1991). As Momeni and Salimi (2023) argued, test takers need to be aware that testing and evaluation should support and reinforce their liberties rather than undermine them.

Test performance is typically treated as a direct reflection of achievement, yet "rarely are there any investigations or discussions as to the sources, causes and reasons that make test items good or bad, easy or difficult" (Shohamy, 2001, p. 4). Davies (2008) notes that tests tend to overlook fairness, ethics, and their own underlying purposes, even though they inevitably embed cultural, historical, and ideological features. Examining tests from a critical perspective can help moderate unequal power distributions among stakeholders and minimize undesirable forms of dominant control over high-stakes tests. As Azizi (2022) emphasized, tests should be designed and evaluated in ways that are perceived as fair by test takers. Ethics in testing thus involves attending to fairness issues and the appropriate use of test data (Taghizadeh & Mazdayasna, 2023).

In Iran, several nationwide English proficiency tests—such as the English Proficiency Test (EPT), the Ministry of Science, Research, and Technology (MSRT) exam, the Test of Language by the Iranian Measurement Organization (TOLIMO), and the Ministry of Health Language Exam (MHLE)—serve as high-stakes exams for PhD candidates. In many cases, PhD graduation depends on achieving minimum required scores on these tests. The consequences of such exams can be severe, including frustration, depression, loss of motivation and self-esteem, negative psychological impact, and delayed graduation (Farah, 2017; Marandi et al., 2020). Addressing these problems requires moving toward more ethical and democratic approaches to testing in Iran and in similar contexts, especially where traditional testing still predominates. In this regard, checklists, as systematic tools for gathering data (McGrath, 2002), can help evaluators consistently attend to critical features of testing conditions and test content without overlooking important criteria.

A review of research on English proficiency tests in Iran reveals two major problems. The first is lack of validity. Pishghadam et al. (2021) examined several popular English proficiency tests in Iran and reported substantial dissatisfaction among test takers; their findings showed that these tests suffer from validity problems and do not accurately reflect test takers' English proficiency. Similar conclusions have been reached by other researchers (Ahmadi & Darabi Bazvand, 2016; Jamalifar et al., 2021; Keyvanfar & Dadfarma, 2013; Rezaeian et al., 2020a). The second problem is unequal power distribution among test stakeholders. Drawing on CLT, Salehi and Tarjoman (2017) found that students and university professors had no meaningful role in developing or administering the MA entrance exam, despite their desire to influence its timing and content. This indicates that such exams are controlled by educational and political elites, thereby reinforcing unequal power relations. Other studies similarly report inequitable power distribution among stakeholders (Khany & Tarlani-Aliabadi, 2016; Razavipour, 2014; Safari, 2016; Safari & Rashidi, 2018; Tahmasebi & Yamini, 2013). Building on these insights, the present study aims to help test developers make high-stakes tests more fair and democratic by acknowledging the voices and rights of test takers and university professors.

The main goal of the checklist proposed in this study is to assist test developers in critically evaluating Iranian English Proficiency Tests (IEPTs), particularly the EPT, prior to administration. Although various standard checklists exist for test evaluation, to the best of the researchers' knowledge, there is no checklist designed explicitly from a critical perspective to uncover hidden assumptions and embedded ideologies in local high-stakes English proficiency tests. Since tests often neglect fairness and ethics while reflecting cultural, historical, and ideological agendas, analyzing them through a critical lens can help rebalance power relations among stakeholders. The proposed checklist therefore aims to minimize inequalities, eliminate unfair and biased items, enhance test justice, improve the validity of IEPTs, and promote a more equitable distribution of power among test parties.

Shohamy (2001) introduced CLT as an extension of critical applied linguistics (Pennycook, 2001), highlighting that tests operate as instruments of power whose use, impact, and social consequences must be monitored. Within this perspective, language testing has shifted from a solely positivist orientation toward interpretive and critical approaches (Lynch & Shaw, 2005; Mertens, 2005). While positivism and post-

10.57647/jelpp.2025.1802.14

positivism emphasize quantitative methods such as multiple-choice tests (Creswell & Plano Clark, 2007), interpretivism and constructivism emphasize qualitative methods like interviews, interpretation, and participants' lived experiences (Cohen & Manion, 1994). In assessment, CLT takes an interpretive stance, viewing test scores not as neutral measures of ability but as context-bound outcomes that must be interpreted in light of social and political factors (Shohamy, 2001a).

From a democratic assessment perspective, tests should not function as instruments of domination. Shohamy (1997) argues that test power is concentrated in the hands of educational and political authorities who manipulate the educational system, creating a division between oppressors and oppressed. In the Iranian testing system, test takers—whose futures depend on high-stakes exam results—may be emotionally, socially, and economically vulnerable to policymakers whose interests are tied to testing (Taghizadeh & Mazdayasna, 2023). McNamara (1996) similarly notes that political groups are often the most powerful actors in testing, using scores to make decisions that shape curricula and control educational systems. In centralized education systems, authorities determine language policies and thus exert considerable influence over high-stakes testing (Shohamy, 2007; Salehi & Tarjoman, 2017). In such contexts, testers and teachers may be unaware of their rights and may feel powerless to advocate for students (Karami, 2013). Noam (1996) underscores that evaluation practices can support careers and success, but they can also damage individuals' self-perceptions and create unnecessary barriers. Razavipour (2014) therefore argues that power relations between test takers and test makers should be reversible, and test takers should be granted a voice in test development.

Recent work has increasingly focused on the critical dimensions of tests, especially test impact and social consequences. CLT is concerned with domains such as gender, class, and ethnicity and with how language is implicated in these domains (Lynch, 2001). Tests with ideological, social, educational, political, and cultural agendas shape individual life trajectories and may be used strategically by test developers and policymakers. CLT emphasizes that testing processes are inseparable from their sociocultural context; test developers must attend to what they include in tests, how tests are administered, and how different groups—particularly minorities—are affected. As Van Dijk (2008) notes, those who control discourse, including testing discourse, may indirectly control people's minds and actions. CLT thus calls for developing critical strategies to examine the uses and consequences of tests, monitor their power, minimize their detrimental effects, expose misuses, and empower test takers by embedding tests within their broader social, ethical, educational, and political contexts (Shohamy, 2001b).

Central to CLT is the expectation that test takers' rights and needs be addressed fairly and without bias, including protection from favoritism related to gender, race, or other factors. Tests should be designed and evaluated according to students' needs (Azizi, 2022), and CLT seeks to identify sources of unfairness rooted in cultural, ideological, and political aspects of testing. Any source of bias that unfairly affects individuals should be detected and corrected (Momeni & Salimi, 2023).

Among the constructive elements of a test, fairness is of paramount importance. Azizi (2022) views fairness as the foundation of any testing procedure, while Kunnan (2000) conceptualizes fairness as encompassing validity, access, and justice. Fairness aims to reduce test bias and increase opportunities for equality in learning and assessment (Xi, 2010). Habermas (1929), cited in Payne & Barbera, 2010) links justice to equal rights and mutual respect, and McNamara and Ryan (2011) distinguish fairness (test properties and psychometrics) from justice (social outcomes, value implications, and interpretations). A fair test is essentially free from bias. Elder (1997) defines test bias studies as efforts to identify and reduce the effects of confounding variables on test scores by modifying the test, and bias can also arise when test content is tailored to a specific group (Rezai et al., 2022). Reducing bias can enhance fairness, which Van de Vijver and Tanzer (2004) describe in terms of construct, method, and item bias.

Fairness and validity are closely interconnected. Willingham (1999) famously stated that "improved validity is the road to fairness" (p. 221). Validity involves gathering evidence to support the correctness of score interpretations (Messick, 1989), and includes face, content, predictive, concurrent, and construct validity. Tests are invalid when they are not used for their intended purpose. In Iran, many high-stakes tests are argued to lack fairness because they lack construct validity (Ahmadi & Darabi Bazvand, 2016; Yaghoobi, 2016). Pishghadam et al. (2021) found substantial dissatisfaction with popular English proficiency tests among PhD students and concluded that these tests lack validity and fail to assess English proficiency accurately, resulting in negative washback. They suggested that localized cultural norms be taken into account in test development. Studies also highlight the importance of standardized testing conditions, including environment, facilities, level of difficulty, preparation, and clear guidance (Coleman, 1998; Roozbehani, 2006).

Research indicates that incorporating test takers' perspectives enhances perceptions of fairness. Doosti and Safa (2021) found that test takers who were

consulted felt more fairly judged and wanted feedback on their strengths and weaknesses. Tofighi and Safa (2023) argued that instructors should be trained in fair assessment procedures to reduce score pollution. Rahimi et al. (2014), in an experimental study on fairness in IELTS and TOEFL iBT, concluded that test preparation classes can have positive washback effects on test interpretations and outcomes.

Consequences are also a key dimension of validity (Messick, 1989). Wang (2016) showed that poor performance on high-stakes tests in the Korean context can place high-achieving students at risk of severe psychological distress, including suicidal ideation. Similarly, Davis (2006) reported that unfair and biased testing leads to dissatisfaction, depression, and demotivation. In Iran, Safa and Sheykholmoluki (2023) highlighted how the Iranian University Entrance Examination (IUEE) can unjustly affect students and parents socially, educationally, and economically, with lasting negative consequences, and called for major reforms to the exam or its governing policies. Conversely, Dornyei and Ushioda (2011) argued that positive test experiences can enhance motivation, achievement, and satisfaction. Rezaeian et al. (2020b) found positive relationships between educational and psychological consequences of the EPT exam, underscoring how affective factors such as motivation and self-confidence are tied to educational outcomes. Nevertheless, test developers can reduce harmful consequences by designing fairer high-stakes tests.

To ensure fairness, one strategy is to examine Differential Item Functioning (DIF). DIF occurs when test takers from different groups but with the same ability level perform differently due to group membership (e.g., ethnicity, gender, religion; Camilli & Shepard, 1994). In their empirical study, Semiyari and Ahangari (2022) found that science majors outperformed humanities majors on the MSRT exam, indicating the presence of DIF linked to subject field. Ahmadi and Darabi Bazvand (2016) documented gender-related DIF in the PhD Entrance Exam of TEFL (PEET) using both Item Response Theory and Logistic Regression.

Socio-economic status is another important factor in language testing. Learners from higher socio-economic backgrounds often have greater opportunities for involvement, engagement, and development, as well as better facilities and instruction (Nikolov & Djigunovic, 2011). Zwick (2006) notes that minorities may perform poorly due to limited educational opportunities and resources. Moses and Nanna (2007), in a U.S. study, found that systematic differences in high-stakes test performance between groups can reinforce existing educational inequalities and harm low-income and minority students, arguing that testing should be used to benefit the least advantaged rather than to harm them.

Cultural and political forces also shape high-stakes, large-scale tests. Policies should be developed in consultation with all stakeholders, especially those most affected (Deygers, 2017). Rezaeian et al. (2020a) showed that examinees' contributions across EPT events can enhance the validity of the exam. Likewise, Choi (2016) found in the Korean context that students had limited knowledge about test content and their rights, leading to tension and confusion.

Kunnan (2004) identified five critical features of tests: validity, absence of bias, equal access for testing and learning, test administration, and social consequences. Access includes personal, educational, financial, material, and geographical dimensions. In some regions of Iran, equal access is not available (Ahmadi et al., 2018; Ghafari et al., 2022; Khosravipour et al., 2023; Nazari et al., 2022). Test takers need familiarity with test content, materials, equipment, procedures, and testing conditions. Accessibility entails nearby test centers, affordable exams, and opportunities for preparation. In the case of EPT, MSRT, MHLE, and TOLIMO, some test takers must travel long distances, experiencing fatigue and reduced concentration. External factors such as lighting, temperature, ventilation, proctor behavior, and consistency across test centers also influence fairness (Kunnan, 2000).

Empirical research has demonstrated the applicability of CLT. Safari and Rashidi (2018) examined three major high-stakes tests in Iran (National University Entrance Exam and MA and PhD entrance exams) and found strong evidence of dominant policies, recommending democratic assessment to minimize negative consequences and promote justice. Khany and Tarlani-Aliabadi (2016) argued that predetermined educational policies render test takers passive, with little opportunity to express their voices. Safari (2016) analyzed the Sampad Entrance Exam (SEE) and the Nationwide University Entrance Exam (NUEE), revealing marginalized test takers, inequitable power distributions, and hidden agendas. She advocated dialogue among stakeholders, multiple assessment sources, and the development of critical thinking skills, warning against the hegemony of testing (Stoddart, 2007). Salehi and Tarjoman (2017) similarly found that TEFL MA teachers had little power in test administration and that tests were largely controlled by educational and political authorities. Although 80% of students and professors expressed willingness to participate in test preparation, authorities resisted their involvement. These exams also tended to emphasize memorization and comprehension

rather than higher-level skills, suggesting the need to assess all language skills rather than focusing narrowly on reading, grammar, and vocabulary.

Stakeholder consultation is thus crucial. Rea-Dickins (1997) highlights the importance of taking stakeholders' views seriously, and Shohamy (2001a) uses CLT to examine the extent of stakeholder participation in testing processes. Tahmasebi and Yamini (2013) studied the development and administration of IUEE and found that power is concentrated in political and educational authorities, with test takers and teachers having little control over exam content or administration. This unequal distribution of power leads to serious consequences and the neglect of some participants' rights. They recommend more democratic testing practices that better serve all stakeholders.

Transparency in testing is also central to CLT. Auerbach (1995) argues that individuals should know about test content, test developers, and evaluation procedures, and Shohamy (2001a) similarly stresses that test takers need to understand the uses of tests. Tahmasebi and Yamini (2013) reported that a large proportion of IUEE participants did not know who developed the test or what sources were used, which runs counter to equal power distribution. Pennycook (2001) suggests that closer relationships between test takers and test makers can reduce decision-making problems. Using CLT, Tahmasebi and Yamini (2013) further highlighted the importance of addressing power distribution, long-term consequences of tests, and their psychological effects.

CLT encourages pupils to be active and critical respondents (Shohamy, 2001a). Empowering students and members of society to think critically is therefore essential. Critical thinking leads stakeholders to reflect on, analyze, and question existing educational policies and practices (Safari, 2016; Safari & Rashidi, 2018). In language testing, democratic assessment can protect individuals from unethical practices (Shohamy, 2001b). Lynch and Shaw (2005) note that ethical issues in assessment are closely tied to power and its potential abuse. Given that admission and graduation often depend on achieving cut-off scores on high-stakes exams such as EPT, MSRT, and TOLIMO (Keyvanfar & Dadfarma, 2013), it is hoped that developing the proposed checklist will help scholars and practitioners reduce the detrimental consequences of such tests and promote fairer treatment of all stakeholders at individual and societal levels.

Method

Participants

The participants of this study were 350 individuals. They consisted of 263 (75.1%) English and non-English PhD candidates (195 females and 155 males), and 87 university professors (24.9%) specialized in different majors. Among the participants, 131 (37.4%) majored in English and 219 (62.6%) majored in non-English fields while 21 of them were experts in CLT with more than eight years of experience. The workplace of 146 (41.7%) participants was university and 204 (58.3%) worked outside the university. The age of the participants in this research ranged from 27 to 58 years, with an average of 35.30 and a standard deviation of 6.801. The participants' first language was Persian. The data was gathered by both on-line and manual questionnaires based on convenience and snowball sampling. The participants were from Islamic Azad University of Gorgan, Qaemshahr, and Mashhad branches, Golestan University, and Tabaran Institute of Higher Education. All the participants were asked to share the link with their friends and colleagues through social media, messaging applications, e-mail, etc.

Instrument

The primary instrument utilized in this study was an evaluative researcher-made questionnaire which was specifically designed to identify and prioritize the essential components involved in constructing high-stakes English proficiency tests (EPT) in Iran. The need for this instrument arose from the lack of a comprehensive, validated checklist aligned with the principles of critical language testing (CLT) and suitable for localized use in Iranian universities by test developers, policymakers, and researchers.

The initial item pool consisted of 215 items, generated through an in-depth review of literature in language testing, CLT, and high-stakes assessment practices. Items were selected based on their theoretical significance, relevance to test quality, and frequency of discussion in prior research. Following expert review, pilot testing, and iterative revisions, the items that were vague, redundant, negatively phrased, or too complex were removed or refined. Ultimately, 66 items were retained based on their psychometric soundness, participant feedback, and expert validation.

To ensure clarity, accuracy, and comprehensibility among the participants from diverse academic backgrounds, the questionnaire was developed in both English and Persian. It was validated in English and Persian language. The questionnaire was translated from English to Persian by a professional translator and then translated back to Persian by another translator familiar with the subject to compare and fix them; then, it was

reviewed by bilingual experts to ensure semantic and conceptual equivalence. Then pilot study was done in Persian language for a small group to gather feedbacks and finally reliability, internal consistency, content and construct validity were assessed. Administering the questionnaire in Persian aimed to eliminate potential language-related misunderstandings, particularly among non-English majors or faculty members less confident in reading academic English. This bilingual approach increased the reliability of responses and allowed participants to express their views more accurately and comfortably.

The finalized questionnaire had two sections. The first section asked for demographic information of the participants and the second section included 66 items, categorized into six components, which were established through exploratory and confirmatory factor analyses including: Content and Validation, Performance, Test Developers, Administration, Distractors, and Scores.

The questionnaire was based on five-point Likert Scale and ranging from strongly disagree to strongly agree. Twenty-one English and non-English PhD candidates and six university professors from diverse universities that were mentioned earlier were interviewed in Persian language (semi-structured interviews). Using Persian language makes them feel comfortable and calm when being interviewed to express their ideas and thoughts. Although there were some pre-prepared questions, the questions were open-ended and the interviewees were asked to elaborate on the suggested issues. The purpose of the interview was to allow the participants to raise points that the researcher had not previously considered. Although the number of participants in the interview was limited, several valuable points were obtained.

Procedure

To fulfill the necessary requirements of this study, there were different steps to follow. As Dornyei (2007) declared, in developing a questionnaire, initial item pool is very essential. Therefore, the questionnaire was generated with 215 items. This was the first step; the items were developed based on the critical review of large volume of related literatures. The researchers tried to use simple and clear words and avoided negative, long and double-barreled words. All of the items were carefully reviewed and examined. After consulting with the experts in testing and CLT, some items were deleted and some were modified in accordance with their recommendations. Then, in the second step and to make sure about the appropriateness of the items of the first draft, a semi-structured interview was piloted with English and non-English PhD candidates and university professors and many beneficial comments were obtained. For the convenience of the interviewees in expressing their ideas, the sessions were not audio-recorded; instead, their comments were documented on writing. Finally, 66 items remained. Afterwards, for those who were not comfortable with English version of the questionnaire, the researchers translated the questionnaire into Persian language. To ensure the accuracy of the translation, the Persian version was double-checked with experts and was validated. Then, the items were grouped by their categories and classified under their corresponding domain.

Subsequently, to pilot the study and following Dornyei (2007), a 5-point Likert scale was used which ranged from “strongly disagree” to “strongly agree”. For initial piloting, it was administered to 35 PhD candidates and many beneficial hints were obtained. Some new ideas and suggestions were mentioned by the participants that were worth considering when deciding which questionnaire items to add, modify, or delete. At the end, the targeted participants cooperated in filling out the final version of questionnaire. The tentative checklist was distributed to 263 English and non-English PhD candidates and 87 university professors of Gorgan, Mashhad and Qaemshahr. However, the participants in different groups of Telegram and Whatsapp (special groups for PhD students and professors in different majors) were given a link of the questionnaire in Google Docs and some of the data was gathered in this way. For 66 items and 350 participants the estimate of reliability was 0.99 which according to Barker et al., (1994), when it is above .70, the instrument has admissible estimation to use.

Design

This study employed a mixed-methods design, combining both quantitative and qualitative approaches. The quantitative phase involved administering a researcher-developed questionnaire to a large sample, while the qualitative phase consisted of semi-structured interviews with a subset of participants to gain deeper insights and enrich the findings. This design allowed for a comprehensive examination of the research by combining statistical analysis with the participants’ perspectives.

Results

At the first step of the data analysis, descriptive statistics for the collected data were examined. Afterwards, the normality of distribution was assessed using skewness and kurtosis values which confirmed that the data followed a normal distribution. To evaluate construct validity, exploratory factor analysis (EFA) was

conducted using SPSS (version 23). Prior to EFA, the Kaiser-Meyer-Olkin (KMO) was computed to ensure sample adequacy. Thereupon, principal component analysis (PCA) with orthogonal varimax rotation was used for component extraction. Then, Cronbach's alpha estimations were used for the selected components, yielding a high reliability score of 0.99, indicating excellent internal consistency. Additionally, to find the probable problematic items, correlation analysis was performed (Field, 2009) and after all 66 items were remained. In the second phase, confirmatory factor analysis (CFA) for testing the model fitness and verifying the appropriateness of the developed questionnaire were performed utilizing AMOSS 22 software.

From 350 participants in this research, 44.3% were men and 55.7% were women, 75.1% were PhD candidates and 24.9% were PhD holders, 37.4% were in the field of English and 62.6% in non-English fields, 41.7% of them worked in university and 58.3% outside the university, Table 1 represents the frequency of demographic variables. Prior to conducting exploratory factor analysis and to ensure the appropriateness of the data for using this method, Keyser, Mayer and O'Klin (KMO) adequacy test and Bartlett's test were performed. The KMO value was 0.985, indicating the appropriateness of the sample size for the questionnaire (Field, 2009). Bartlett's level of significance in the questionnaire was 0.000, demonstrating sufficient correlation between the variables and the suitability of the data for factor analysis. The KMO value was 0.985, indicating that the sample size was appropriate for the questionnaire (Field, 2009). Bartlett's test of significance was 0.000, demonstrating sufficient correlations among the variables and confirming the suitability of the data for factor analysis. Bartlett's test of sphericity showed that the results are statistically significant for the patterned correlations between the items, $\chi 2\ (2145) = 25942.697$, $P \leq .001$ (Table 2).

Before conducting the main analysis and extracting the components, the researchers needed to use Kaiser's criterion which was factors with eigenvalue of 1 or higher. The results confirmed that six components were above the point of inflexion (more than 1). After rotation for components 1 to 6, the results indicated that the eigenvalue were 27.112, 5.084, 4.733, 4.668, 3.959, and 2.480, respectively. According to Field (2009), the minimum value of factor loading was determined to be 0.4. The initial and rotated eigenvalues are presented in Table 3.

The results in Table 3 show that six components have a specific value above 1, which can be used to obtain 72.78 of the results. According to the results, all of six components could be extracted. These six extracted components were used for factor loadings after rotation. The process of factor loadings formed the final phase in principal component analysis. As Field (2009) declared, factor loadings "were a gauge of substantive importance of a given variable to a given factor" (p. 644).

Table 1. Frequency of Demographic Variables (N=350)

Variable	Category	Frequency	Percent
Sex	Male	155	44.3%
	Female	195	55.7%
Level of Education	PhD Student	263	75.1%
	PhD	87	24.9%
Major	English	131	37.4%
	Other Majors	219	62.6%
Working Place	University	146	41.7%
	Other	204	58.3%

Note. Percentages are rounded to one decimal place

Table 2. KMO and Bartlett's Test

Test	Statistic	df	Significance(Sig.)
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	.985	—	—
Bartlett's Test of Sphericity	Approx. Chi-Square	2145	.000
	25942.697		

Notes: df' stands for degrees of freedom. Sig. represents significance level. The dash (—) indicates data not applicable

Table 3. Eigenvalues Before and After Rotation

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	40.647	61.586	61.586	40.647	61.586	61.586	27.112	40.948	40.948
2	2.630	3.986	65.572	2.630	3.986	65.572	5.084	7.582	48.530
3	1.320	2.001	67.573	1.320	2.001	67.573	4.733	7.142	55.672
4	1.259	1.908	69.480	1.259	1.908	69.480	4.668	7.088	62.760
5	1.143	1.732	71.213	1.143	1.732	71.213	3.959	6.599	69.359
6	1.036	1.569	72.782	1.036	1.569	72.782	2.480	3.471	72.830

Extraction Method: Principal Component Analysis

By consulting experts and carefully examining the questions within each component, the appropriate label for that component was determined. The questionnaire containing 66 items converted into six components named as Content and validation, Performance, Test Developers, Administration, Distractors, and Scores. The six components and their related questions are provided in Appendix 1.

To compute the internal consistency of the obtained components and indicating that all items reflect their components, Cronbach's alpha was used for each component. The highest reliability value related to Content and validation was 0.991 and the lowest reliability related to Performance was 0.761.

Table 4. Reliability Statistics of Components

Component	Cronbach's Alpha	N of Items
Content and validation	.991	51
Performance	.761	3
Test Developers	.873	4
Administration	.839	5
Distractors	.853	2

Reliability was not calculated for scores since it included only one question. According to Barker et al. (1994), this estimation represented an acceptable internal consistency of items and components (Table 4).

To verify the appropriateness of the developed model and fitness of the data, the next step, confirmatory factor analysis (CFA) was conducted. To this end, AMOS 22 software was used. Some indices were evaluated to examine the fitness of model, including the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), Incremental Fit Index (IFI), Tucker-Lewis Index (TLI), and Discrepancy Divided by Degree of Freedom (CMIN/DF). According to Tseng and Schmitt (2008), for a model to demonstrate good fit, the $\chi 2/df$ ratio should be less than 3. The obtained value was 2.145, that is acceptable. In the tentative model, the $\chi 2$ value was 4287.186 with 1999 degrees of freedom. In addition, the results showed that CFI was .910, IFI was .910, and TLI was .906. According to Hu and Bentler (1999), when these values are above .90, the results are acceptable. Moreover, according to Hu and Bentler, the RMSEA should be close to .06. The results in our table showed an RMSEA of .057,

indicating acceptable model fit. Table 5 shows that all goodness-of-fit indices are within the acceptable range; so, the final model demonstrated a good fit to the data. In the figure below, the standardized structural equation model of the researcher-made questionnaire is displayed.

After factor analysis and to assess the correlation between six-factor model, the following data were obtained. It shows that there is a significant two-to-two linear correlation between all the factors (Table 6).

Table 5. Model Fit of Quality of Life

Fit Index	Value	Criterion	Desirability
RMSEA	.057	< 0.08	Desirable
IFI	.910	> 0.09	Desirable
TLI	.906	> 0.09	Desirable
CFI	.910	> 0.09	Desirable
CMIN/DF	2.145	< 0.03	Desirable

Note. RMSEA = Root Mean Square Error of Approximation; IFI = Incremental Fit Index; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; CMIN/DF = Minimum Discrepancy/ Degrees of Freedom

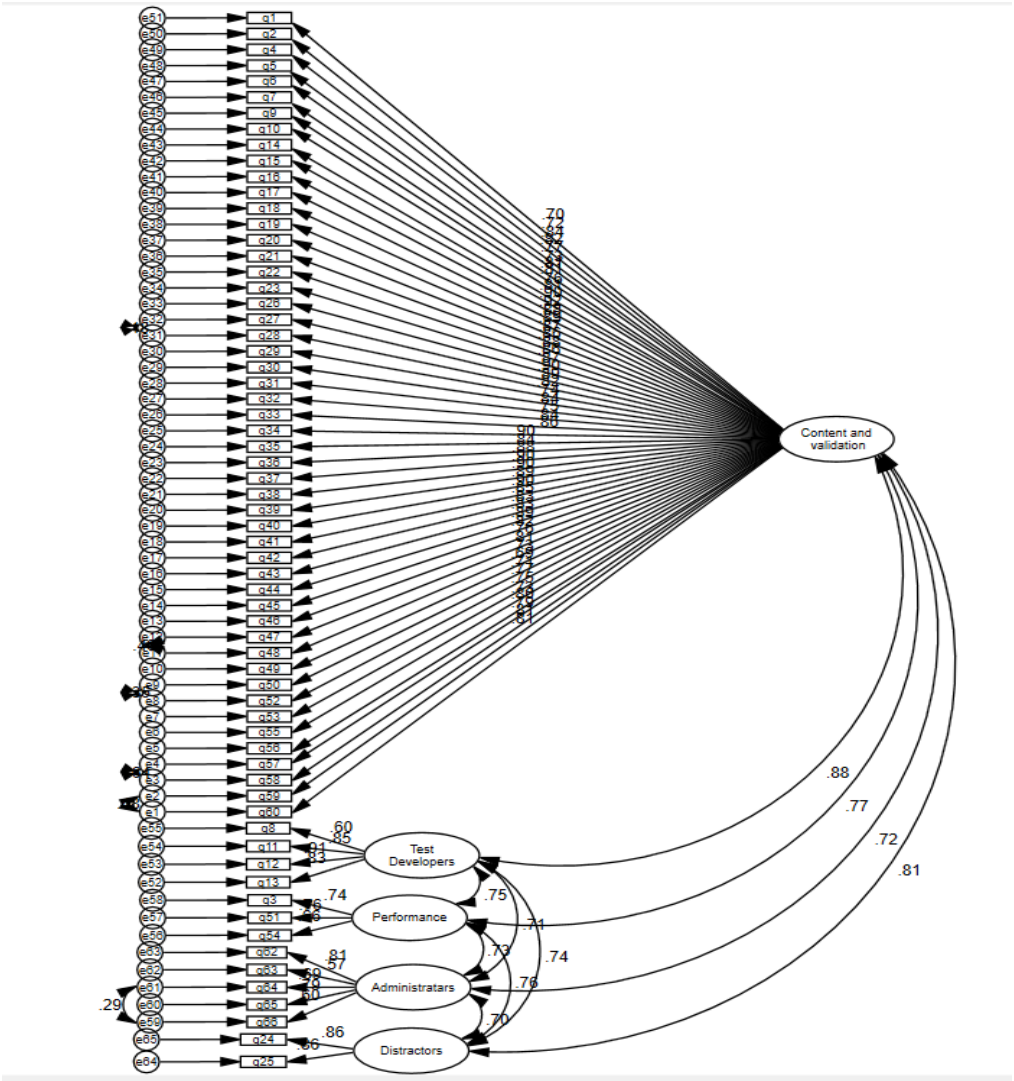

Figure 1. The Standardized Model of the Structural Equations of the Researcher-Made Questionnaire

Table 6. Correlations Between Factors

Factor	Content & validation	Test Developers	Performance	Administration	Distractors	Scores
Content & validation	1.000					
Sig. (2-tailed)						
Test Developers	.691	1.000				
Sig. (2-tailed)	.000					
Performance	.833	.622	1.000			
Sig. (2-tailed)	.000	.000				
Administration	.589	.565	.567	1.000		
Sig. (2-tailed)	.000	.000	.000			
Distractors	.411	.475	.370	.563	1.000	
Sig. (2-tailed)	.000	.000	.000	.000		
Scores	.411	.475	.370	.563	1.000	1.000
Sig. (2-tailed)	.000	.000	.000	.000	.000	

Note. Pearson correlations are reported; Sig. indicates the significance (2-tailed)

Discussion

This study made an effort to develop and validate a critical checklist for high-stakes English proficiency tests and specifically EPT exam. To this aim and based on the critical review of the related literatures, 215 items were identified. After consulting with experts and semi-structured interviews some of the items were modified or deleted, and, eventually, a 66-item checklist was developed. The tentative questionnaire was distributed among PhD candidates and university professors of Gorgan, Mashhad and Qaemshahr. After confirming the normality of distribution of the data and to pursue the research, exploratory factor analysis (EFA) for construct validity evaluation, principle component analysis for component extraction and confirmatory factor analysis (CFA) for verifying the appropriateness of the developed questionnaire were performed. The results of EFA, PCA, CFA, and reliability estimates confirmed the reliability and validity of the tentative new developed instrument. Some indices were evaluated to examine the fitness of model and all indices were found to be within the acceptable range, so the final model demonstrated a good fit to the data. After factor analysis, the researchers assessed the correlation between six-factor model, and the results showed that there was a significant two-to-two linear correlation between all the factors.

The critically developed items of the checklist are in line with Friere's (1970) problem-posing plan of education. In his plan, he stressed the partnership and mutual relationships between learner and instructor or test taker and test developers. Besides, he alerts the probable social, political, and economic dominations imposed on learners and tried to draw attentions to equality, fairness and rights of learners. The focus of his form of education was on critical consciousness considering liberties "so that through transforming action they can create a new situation, one which makes possible the pursuit of fuller humanity" (p.47). Items number 2, 12, 13, 25, 26, 29, 44, 46, 54 reflect his concept of problem-posing education. In addition, the items of the checklist were developed in line with Shohamy's (2001) critical language testing principles in which social, educational, political, cultural, and ideological agendas matter. She emphasized exploring the needs and rights of test takers fairly. Highlighting the importance of the use of tests, she also considered the amount of participation and influence of different stakeholders in tests. Moreover, she was concerned about equality and social justice in tests development in which test takers should have an active role apart from any favoritism or bias. Items number 7, 8, 9, 20, 23, 24, 27, 28, 29, 31, 46, 53, 54 reflect her principles of critical language testing. Furthermore, in line with Lynch (2001), in developing items, ethical considerations, validity, fairness, ethnicity, class, and gender were taken into consideration. Items number 9, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 45, 46, 50, 51, 56 reflect his principles of language test evaluation. Besides, due to relevance of ethics and power relations, Foucault's (1982) theory of ethical power relations were noted in which he mentioned the importance of the amount of stakeholders' control

on assessment policies. Items 3, 4, 5, 6, 8, 9, 18, 20-28, 42, 43, 46, 54, 59, 63, 66 reflect his ideas about knowledge and power. Developing this checklist is in line with Messick's (1989) notion in which he believed that ethical way of assessment should be applied focusing on tests consequences. Further, for developing items and in line with Willingham and Cole (1997) and Kunnan (2000), tests should be administered in equal situations and equal scoring procedure for all test takers.

Developing a critical checklist for high-stakes English proficiency tests in Iran is a breakthrough in the field of language education. Since in Iran the graduation of PhD candidates is dependent on passing one of English proficiency exams, they can have very remarkable consequences as was discussed formerly. Therefore, considering ethical and fairness issues in designing such tests must be at the center of attention. In developing this checklist, different aspects of test development process were considered. Careful studies of related literatures attracted the authors' attention to most problematic areas based on which the items of checklist were developed. Considering the problems, the tentative checklist is developed against injustice in favor of minorities which helps test developers design and develop tests in more democratic ways comparing to before. Although, the literature regarding this topic was very limited, the researchers utilized trusted methods of construct development and validation processes affirmed by Dornyei (2007).

The present study explored 66 items for evaluating English high stakes tests in Iran from the critical point of view. This researchers-made checklist was developed based on critical language testing principles in which it has been attempted to provide test takers with impartial, just and fair tests and testing circumstances all over the country. By making the checklist accessible to test developers, they will have the opportunity to appraise the EPT exam from the critical point of view before tests are being developed. This study explored six components that researchers believed could cover different aspects of test development process against unfair and unjust behavior toward test takers. In fact, this kind of critical checklist makes test developers alert to possible inequalities and notable consequences of EPT exam on test takers and their parents. Using the tentative checklist moderates power distribution among test parties while voices and rights of test takers can be heard and observed.

The achievements of this study and the results of the interviews with test takers revealed that test takers and even professors were willing to take part in the process of test development and administration. They should have equal geographical access to exam location and educational access to preparation classes. Moreover, they believed that selective decisions based on only one multiple choice exam is not justified, and that interpretative techniques should be added. Test takers wished to be allowed to share their comments and criticisms on EPT exam with test developers. It would be better if the EPT administration site were also close to test takers since many of them have to travel from other cities to take the exam. The behavior of test personnel is important due to its effect on test takers. Moreover, external factors of the EPT exam, such as lighting, temperature, etc. should be monitored across all locations nationwide. Furthermore, the EPT should be designed based on a needs analysis and the academic majors of the test takers. Although this exam has notable negative consequences due to test takers' failures, many reported that even after passing this exam, they still were unable to apply their skills in real life situations, as their studies focused primarily on passing the test rather than practical application.

Therefore, as McNamara (1996) stated, in testing process, the power is in the hands of the political groups who impose curricula and make decisions based on the test scores. Shohamy (2007) further noted that authorities assign the educational policy. Nevertheless, although policymakers are dominant in Iran (Taghizadeh & Mazdayasna, 2023), they should have considered implementing significant changes to their policies (Safa & Sheykholmoluki, 2023).

While test developers are the primary users of this checklist, it can be also valuable for teachers who regularly work with tests and policy makers whose decisions and policies impact test takers. However, this checklist has several limitations that should be considered. First, the research was carried out in the cities of Gorgan, Qaemshahr, and Mashhad with 350 respondents. Further studies are highly encouraged to replicate this research in other cities with a larger and more diverse group of participants to enhance the validity of the results. Second, the checklist

developed in this study was specifically for the EPT exam. Although it can serve as a valuable tool for test developers, policy makers, and teachers who work regularly with assessments, further research is needed to adapt and validate the checklist for other English proficiency exams such as MSRT or MHLE and so on. Finally, it is also recommended that more items be developed and added to this checklist which can cover more aspects of test development process from the critical point of view.

Authors Contribution
All authors have contributed equally to prepare the paper.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflict of interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmadi, A., & Bazvand, A. D. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research 4*(1), 63-82

Ahmadi, S., hassani, M., & Mousavi, M. (2018). Institutional Differentiation and Inequality in Iran Education System (A Case Study of Second- Cycle High Schools in Urmia (. *Journal of Applied Sociology*, *29*(2), 169-190.
DOI: https://doi.org/10.22108/jas.2017.102412.1066

Ashraf, H., & Khosravani, M. (2021). On the effect of perceived test impact on test anxiety and attainment of PhD candidates: Studying English Proficiency Test in Islamic Azad University. *International Journal of Linguistics, Literature & Translation*, *4*(1), 10–16.
DOI: https://doi.org/10.32996/ijllt.2021.4.1.2

Auerbach, E. R. (1995). *The teacher as researcher: A critical perspective*. TESOL Quarterly, *29*(3), 443-455.

Azizi, Z. (2022). Fairness in assessment practices in online education: Iranian university English teachers' perceptions. *Language Testing in Asia*, *12*(1), 1-17.

Bachman, L. F. (2013). Ongoing challenges in language assessment. In B. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 4, pp. 1586–1604). Wiley-Blackwell.

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or gate- openers? In B. Spolsky & F. M. Hult (Eds.). *The Handbook of Educational Linguistics* (pp. 456-468). Oxford, UK: Blackwell Publishing.

Baker, D. (1989). *Language testing: A critical survey and practical guide*. London: Edward Arnold.

Barker, C., Pistrang, N., & Elliot, R. (1994). *Research methods in clinical and counselling psychology*. Oxford: John Wiley & Sons.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F.J.R.C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes, and prior knowledge*
(pp. 3-29). Dordrecht, Netherlands: Kluwer

Bourdieu, P. (1991). *Language and symbolic power*. Cambridge: Harvard University Press.

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

Choi, In-hye. (2016). Test-takers' Perceptions of Test Fairness: A washback study on test information given prior to a high-stakes writing test. *Foreign Language Education Research*, *19*(2), 1-18.

Cohen, L., & Manion, L. (1994). *Research Methods in Education* (4th ed.). London: Routledge.

Coleman, A. L. (1998). Excellence and equity in education: High standards for high-stakes tests. *Virginia Journal of Social & the Law*, *6*(10), 81-114.

Creswell, J.W., & Plano Clark, V.L. (2007). *Designing and Conducting Mixed Methods Research*. London: Sage.

Davis, A. (2006). High stakes testing and the structure of the mind: A reply to Randall Curran. *Journal of Philosophy of Education*, *40*(1), 1–16.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing 25*(3) 327 – 348.

Deygers, B. (2017). Just testing, Applying theories of justice to high-stakes language tests. *International Journal of Applied Linguistics 168*(2), 143–162.

Doosti, M., & Ahmadi Safa, M. (2021). Fairness in Oral Language Assessment: Training Raters and Considering Examinees' Expectations. *International Journal of Language Testing*, *11*(2),64-90.

Dornyei, Z. (2007). *Research methods in applied linguistics*. New York: Oxford University Press.

Dörnyei, Z. (2010). *Questionnaires in second language research: Construction, administration, Processing* (2nd ed.). Routledge.

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and Researching Motivation*. 2nd ed. Harlow: Longman

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, *14*(3), 261-277.

Farah, M. (2017). Accountability Issues and High Stakes Standardized Assessment: Retrieved from.
DOI: https://doi.org/doi:10.7282/T3W95CZ8

Freire, P. (1970). *Pedagogy of the oppressed*. New York: The Continuum International Publishing.

Freire, P. (1985). *The Politics of Education*. Hadley, MA: Bergin and Garvey

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.

Foucault, M. (1982). The subject and power. In H. L. Dreyfus & P. Rabinow, (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics* (pp. 208–26). Brighton: Harvester Press.

Ghafari Esmaili, S.A., Hassani, M., & Ghalavandi, H. (2022). An Analysis of the Development Level of Educational Indicators as a Factor for Assessing Inequalities and Achieving Sustainable Educational Justice: A Case Study, Education Areas of Mazandaran Province. *Journal of Educational Planning Studies*, *11*(21), 33-52.
DOI: https://doi.org/10.22080/eps.2022.22184.2062

Habermas, J. (1929). *A Dictionary of Cultural and Critical Theory*. 2nd edition. 328-322

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. *A Multidisciplinary Journal*, *6*(1), 1-55

Jamalifar, G., Salehi, H., Tabatabaei, O., & Jafarigoha, M. (2021). Washback Effect of EPT on Iranian PhD Candidates' Attitudes Toward Learning Materials. *International Journal of Foreign Language Teaching and Research, 9* (39), 11-26.

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.

Khany, R., & Tarlani-Aliabadi, H. (2016). Studying power relations in an academic setting: Teachers' and students' perceptions of EAP classes in Iran. *Journal of English for Academic Purposes, 21*, 72–85.

Khosravipour, Z., Mehralizadeh, Y., Darafsh, H., Rahimidoost, G., & Karami, M. (2023). Analysis of Educational Inequality Policies in Six Development Programs in Iran. *Iranian Journal of Public Policy*, *9*(3), 165-181.
DOI: https://doi.org/10.22059/jppolicy.2023.95729

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium*, Orlando, Florida (pp. 1-14). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M., Milanovic & C., Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge: Cambridge University Press.

Keyvanfar, A., & Dadfarma, N. (2013). Bias and Differential Item Functioning in Post-Graduate English Proficiency Tests in Iran. *Middle-East Journal of Scientific Research, 17* (2), 205-212

Lynch B. (2001), Rethinking assessment from a critical perspective. *Language Testing*, *18*(4), 351-372.

Lynch, B., & Shaw, P. (2005). Portfolios, power, and ethics. *TESOL Quarterly*, *39*(2), 263-298.

Marandi, S. S., Tajik, L., & Zohali, L. (2020). On the Construct Validity of the Iranian Ministry of Health Language Exam (MHLE,). *Journal of Language Horizons, Alzahra University,4(2),* Autumn – Winter (Biannual – Serial No. 8

McGrath, I. (2002). *Materials evaluation and design for language teaching*. *Edinburgh*: Edinburgh University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T., & Ryan, K. (2011). Fairness versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*, 161-78.

Mertens, D.M. (2005). *Research Methods in Education and Psychology: Integrating Diversity with Quantitative and Qualitative Approaches*. (2nd ed.) Thousand Oaks: SAGE.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: National Council on Measurement in Education and American Council on Education.

Moses, M. S., & Nanna, M. J. (2007). The Testing Culture and the Persistence of High Stakes Testing Reforms. *Education and Culture*, *23*(1), 55-72.

Mohammadi Roozbehani, K. (2006). Introduction to standardization of exams: Investigating the conditions of holding national entrance exams to universities by measuring the satisfaction of participants. *Research and Planning in Higher Education, 3*(41), 1-12.

Momeni, A., & Salimi, E. A. (2023). Moving toward a Democratic Assessment Framework: Iranian EFL Teachers' Critical Language Assessment Literacy. *International Journal of Language Testing*, *13*(2), 13-37.

Nazari, F., Pirootiaghdam, M., & Zovko, M. (2022). Educational inequalities in Iran based on the viewpoints of educational experts and qualified high school teachers. *Distinctio, 1* (2), 73-93.
DOI: https://doi.org/10.56550/d.1.2.3

Nevo, D. (1996). *School based evaluation*. Oxford: Pergamon University Press

Nikolov, M., & Mihaljević Djigunović, J. (2011). All shades of every color: An overview of early teaching and learning of foreign languages. *Annual Review of Applied Linguistics, 31*, 95–119.

Noam, G. (1996). Assessment at a crossroads: Conversation. *Harvard Educational Review, 66*(3), 631–657.

Öztürk Karatas, T., & Okan, Z. (2019). The Power of Language Tests in Turkish Context: A Critical Study. *Journal of Language and Linguistic Studies*, *15*(1), 210-230.

Payne, M., & Barbera, J. R. (Eds.). (2010). *A dictionary of cultural and critical theory*. West Sussex: Blackwell Publishing Ltd.

Pennycook, A. (2001). *Critical applied Linguistics: A critical introduction*. London: Rutledge.

Pishghdam, R., Ebrahimi, S., Shayesteh, S., Tabatabaee Farani S., & Jajarmi, H. (2021). Evaluating and Examining the Problems of Iran's Ministry of Higher Education and Universities' English Language Proficiency Tests and Stakeholders' Language Needs Analysis. *Journal of Foreign Language Research*, *10* (4), 686-705.

Rahimi, F., Bagheri, M. S., Sadighi, F., & Yarmohammadi, L. (2014). Using an Argument-based Approach to Ensure Fairness of High-stakes Tests' Score-based Consequences. *Procedia - Social and Behavioral Sciences 98*, 1461 – 1468

Rezaeian, M., Seyyedrezaei, S. H., Barani, G., & Seyyedrezaei, Z. S. (2020a). Construction and Validation of Educational, Social and Psychological Consequences Questionnaires of EPT as a High-Stakes Test. *International Journal of Language Testing*, *10*(2), 33-70.

Rezaeian, M., Seyyedrezaei, S. H., Barani, G., & Seyyedrezaei, Z. S. (2020b). An Investigation into Iranian Non-English PhD Students' Perceptions Regarding Learning as an Educational Consequence of EPT. *Iranian Journal of Learning and Memory*, *3*(10), 71-80.

Razavipour K. (2014), The Power of Powerless Language Tests: Test Takers' perceptions, *The Iranian EFL Journal*,*10*(2). April Issue.

Rea-Dickins, P. (1997). So why we need relationship with stakeholders in language testing? A view from UK. *Language Testing*, *14*(3), 304–314.

Rezai, A., Namaziandost, E., Miri, M., & Kumar, T. (2022). Demographic biases and assessment fairness in classroom: insights from Iranian university teachers. *Language Testing in Asia*, *12*(8).

Safa. M. A., & Sheykholmoluki. H. (2023). An impact study of the Iranian National University Entrance Exam from students and parents' perspectives. *Language Testing in Asia*, 13, article 40.

Safari, P. (2016). Reconsideration of Language Assessment is a MUST for Democratic Testing in the Educational System of Iran, *Springer, Interchange 47*, 267–296.

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

Salehi, M., & Tarjoman, M. (2017). An investigation of a nationwide exam from a critical language testing perspective, *Cogent Social Sciences*, *3*(1), 1396639

Semiyari S. R., & Ahangari S. (2022). Examining Differential Item Functioning (DIF) For Iranian EFL Test Takers with Different Fields of Study. *Research in English Language Pedagogy*,*10* (1), 169-190

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14(3), 340–349.

Shohamy, E. (2001a). *The power of tests: a critical perspective on the uses of language tests*. London: Longman.

Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373 391.

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education*, *14*(1), 117–130.

Stoddart, M. C. J. (2007). Ideology, hegemony, discourse. A critical review of theories of knowledge and power. *Social Thought & Research*, *28*, 191–225.

Taghizadeh. M., & Mazdayasna. G. (2023). Language assessment course at Iranian state universities: An evaluation of the incorporation of assessment principles into the course content. *Heliyon*, *9*, e12857

Tahmasebi, S., & Yamini, M. (2013). Power Relations among Different Test Parties from the Perspective of Critical Language Assessment. *The Journal of Teaching Language Skills (JTLS)*, *4* (4), winter, Ser. 69/4

Tofighi, S., & Ahmadi Safa, M. (2023). Fairness in Classroom Language Assessment from EFL Teachers' Perspective. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *42*(2), 81-110.

Tseng, W. T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*, 357-400.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *International Journal of Research in Linguistic, Language Teaching and Testing*, *54*(2), 136-143.

Van Dijk, T. A. (2008). *Discourse and power*. Palgrave Macmillan.

Wang, L. C. (2016). The Effect of High-Stakes Testing on Suicidal Ideation of Teenagers with Reference-Dependent Preferences. *Journal of Population Economics*, *29*, 345–364.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Willingham, W. W. (1999). A system view of test fairness. In S. J. Messick (Ed.). *Assessment in higher education: Issues of access, student development, and public policy* (pp. 213-42). Mahwah, NJ: Lawrence Erlbaum

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. DOI: https://doi.org/10.1177/0265532209349465

Yaghoobi, M. (2016). Fairness and Bias in Language Testing. *International Journal of Research in Linguistics, Language Teaching and Testing*, *1*(3), 136- 143

Zwick, R. (2006). Higher education admission testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 647-80). Westport, CT: Praeger Publishers.

APPENDIX 1

Detailed Categories of Questions

Component	Questions	
Content and validation	1	EPT is developed based on the existing linguistic needs of test takers.
	2	EPT is developed based on the existing social needs of test takers.
	3	Test developers set the desired objectives of EPT exam prior to test making process.
	4	After setting objectives, test developers specify what items to include in each section of EPT.
	5	After setting objectives, test developers specify how many items to include in each section of EPT.
	6	Prior to test making, test developers generate construct definition. (meaning of what is intended to be measured)
	7	Test developers observe code of ethics in the process of making the test.
	8	Test developers observe fairness issues in the process of making the test.
	9	Test developers keep in mind the negative washbacks of EPT when developing test.
	10	Items are arranged from the easiest to the most difficult ones.
	11	EPT is developed based on previous empirical research findings.
	12	Reading comprehension texts are selected from real-life contexts.
	13	The topics of reading comprehension sections are relevant to social needs of test takers.
	14	There is local independence among test items (the answer to one item is independent from the next item).
	15	The language in the EPT exam is authentic.
	16	The duration of EPT exam is appropriate.
	17	The instructions in the EPT exam are easy to understand.
	18	EPT test items measure the single intended ability (unidimensionality).
	19	Several passages are included in reading comprehension section to avoid bias in favor or against some test takers.
	20	Modified form of EPT is used for any disabled test taker.
	21	The topics of readings are neutral for all test takers with different knowledge backgrounds.
	22	Items of exam exclude technical terminologies (e.g. geology, biology, etc...).
	23	The content of EPT is checked to avoid the superiority of one especial ethnic group/culture.
	24	The content of the new version of EPT exam is ideologically neutral and unbiased.
	25	The topics of reading comprehension sections contains critical /controversial subjects.
	26	The topics in reading comprehension sections encourage critical thinking in test takers.
	27	The items of the EPT exam are checked for sexist vocabulary
	28	The items of the EPT exam are checked for racist vocabulary.
	29	To increase the fairness, the viewpoints of test takers from the previous EPT exams are applied in the new version.
	30	The previous shortcomings of the EPT exam are eliminated in the newly developed one.
	31	EPT is designed based on inoffensive language and content.
	32	The initial draft of test items is peer-reviewed before piloting.
	33	The problematic items of EPT exam have been revised or omitted.
	34	EPT is piloted with a representative sample prior to administration.
	35	Test makers use appropriate standards for developing EPT exam (e.g. TOEFL).
	36	Item difficulty is calculated for the items in the EPT exam.
	37	Item discrimination is calculated for the items in the EPT exam.
	38	Reliability of EPT is checked.
	39	To ensure the reliability, EPT is piloted with IRT models.
	40	To check the construct validity of the EPT scores, factor analysis is used.

Component	Questions	
Test Developers	41	The Standard Error of Measurement (SEM) confidence intervals are used to minimize the false decision making on test takers.
	42	To determine the fair cut-off score, mean of scores is computed.
	43	The currently developed EPT has been checked with TOEFL/TOEIC exam to confirm there is high correlation between them.
	44	EPT is developed in such a way to be authentic (test takers can use what they learned in real life situations).
	45	To compare the results of the analysis before and after removing problematic items, factor analysis is conducted on the whole test.
	46	Test takers' views and expectations are being considered in test making process.
	47	After piloting, low item facility values are reconsidered.
	48	After piloting, high item facility values are reconsidered.
	49	After piloting, low item discrimination values are reconsidered.
	50	To evaluate test results, Item Response Theory (IRT) is used.
	51	To evaluate test results, Differential Item Functioning (DIF) techniques are used.
	52	Test developers use an equal number of questions for humanities-oriented and science-oriented subjects.
	53	Test developers avoid the effect of their personal characteristics on item writing process (personality, cultural background, social status, etc.).
	54	In the process of developing EPT exam, different groups of people with varied socio-cultural backgrounds take part.
	55	To prevent from test wiseness, test developers consider their idiosyncracies when making EPT.
Performance	56	The average performance of test takers shows that the difficulty level of test items is adequate.
	57	Factor analysis is used to distinguish traits from methods.
	58	The high values of the mean and median indicates that the overall performance of test takers is satisfactory.
Administration	59	Test proctors are trained to behave in a way to avoid negative influences on test takers.
	60	The quality of testing administration site is equally the same all around the country.
	61	The adequacy of external features of test (e.g., light, temperature, ventilation, etc.) is checked for all test takers throughout Iran
	62	EPT administrators are trained in how to administer the exams (even the use of accommodations).
	63	In EPT exam, modified administration procedure is used for any disabled test taker.
Distractors	64	Distractors in the EPT exam are well distributed.
	65	Distractors in the EPT exam are plausible.
Scores	66	For each section of EPT, separate scores are reported to test takers.

Biodata

Shaban Najafi Karimi is an assistant professor at Islamic Azad University, Qaemshahr Branch. His main research interests include teaching language skills and materials development.

Farzaneh Nouri is currently a PhD candidate in TEFL at Islamic Azad University, Qaemshahr Branch, Iran. Her research interest is critical language testing.

10.57647/jelpp.2025.1802.14