RESEARCH

Open Access

EFL teachers' cognition of social and psychological consequences of high-stake national language tests: role of teacher training workshops

Rahmatolah Allahyari[1], Mahmoud Moradi Abbasabady[2], Shamim Akhter[3] and Goudarz Alibakhshi[4*]

*Correspondence:
alibakhshi@atu.ac.ir

[1] Faculty of Psychology and Education, University of Tehran, Tehran, Iran
[2] University of Mazandaran, Babolsar, Iran
[3] School of Languages, Civilisation and Philosophy, University Utara Malaysia, Sintok, Malaysia
[4] Allameh Tabataba'i University, Tehran, Iran

Abstract

Consequential validity, a facet of construct validity, has been extensively explored by educational psychologists and scholars focused on English language assessment. It is widely recognized that English language instructors must possess a thorough understanding of high-stake language tests. However, the body of research concerning EFL teachers' perceptions of high-stake tests is quite limited. This study aims to delve into the cognitions of Iranian EFL teachers regarding the social and psychological ramifications of high-stake English language tests. Additionally, the study investigates the influence of a teacher training workshop on EFL teachers' perceptions of test consequences. The research employs a two-phase quantitative research design. A total of 210 EFL teachers participated in the survey, completing a scale comprising 25 items that delineated their perceptions of test consequences. Furthermore, thirty teachers engaged in a two-session workshop focused on test consequences. Data analysis encompassed both one-sample and paired-sample *t* tests.

The results reveal that EFL teachers possess knowledge regarding certain social and psychological consequences of language tests; however, their awareness of some adverse consequences remains limited. Notably, the outcomes indicate that teacher training workshops have a positive impact on teachers' cognitions of both positive and negative test consequences. The implications of these findings extend to teacher trainers and English language educators, offering practical insights to enhance the effectiveness of their teaching practices.

Keywords: Assessment literacy, Teachers' cognitions, Test consequences, Consequential validity, High-stake tests

Introduction

School and tertiary teachers/instructors bear multiple responsibilities related to assessment, encompassing tasks such as selecting and designing assessment methods, determining grades, providing feedback, administering assessment tasks, and communicating student achievements to various stakeholders including parents, administrators, students, potential employers, and fellow educators (Alibakhshi &

Shahrakipour, 2014; Butler, 2021; Popham, 2011, 2014; Russell & Airasian, 2012; Vogt, et al., 2020; Tajeddin et al., 2022; Traylor, 2013). Extensive research underscores that teachers dedicate a substantial portion of their time to evaluation activities (Bachman, 2014). Consequently, classroom assessments directly influence education and student learning (Alibakhshi & Mezajin, 2013; Earl, 2013; Green, 2014; Kremmel & Harding, 2020; Sultana, 2019; Tsagari, 2017). To align with the goals of twenty-first century education and equip students for lifelong learning, teachers must seamlessly integrate assessments into their teaching methodologies and students' learning processes (Popham, 2014).

The concept of educational assessment has undergone significant evolution over time. Initially, the primary objective of educational evaluation was to acquire the most precise and dependable information feasible (Board, 2003, p. 55). Subsequently, assessment evolved into a mechanism for educating, instructing, and reshaping educational environments (Brown, 2019; Levi & Inbar-Lourie, 2020; Leenknecht et al., 2020). This shift in perspective has brought attention to the consequences, washback, and instructional nature of assessments (Jönsson, 2020). Shohamy (2001a) introduced the concept of critical assessment, introducing an additional layer to school evaluation in her work "The Power of Tests." Assessments are acknowledged as tools with policy implications capable of driving changes in education and society. The notion of critical assessment has expanded the scope of educational assessment beyond classroom settings, encompassing political, cultural, and social contexts (Shohamy, 2001b, 2007, 2009, 2017).

Language evaluation has similarly undergone significant procedural and conceptual changes within the realm of educational assessment. Language testing, originally focused on measurement, has now shifted to incorporate political and social considerations. Language testers now emphasize the political and social functions of language assessments. Messick (1981) suggested that tests related to psychological, social, and political systems should be understood, given their impact on curriculum, ethics, social hierarchies, bureaucracy, politics, and language understanding. Presently, language assessment research delves into test ethics, bias, instructional influence, and the utilization of tests for power and control (Shohamy, 2001a).

Awareness of the social consequences of tests emerged from validity studies, particularly within the work of Samuel Messick (1989b). Messick introduced a validity model that encompassed two levels of viewing validity: the interpretation of test scores and the actual utilization of tests. He recognized that the construct of a test and its administration held "value implications," signifying the social and cultural values that shape interpretations of test results. Furthermore, Messick's validity framework encompassed the tangible consequences of tests within social contexts. This concept is now known as consequential validity or impact (Brown & Abeywickrama, 2010).

Following Messick's influential contributions, other validity paradigms predominantly overlooked social aspects. McNamara and Roever (2006), however, contend that disregarding the social functions of tests hinders progress in language testing and advocates for an expanded perspective beyond traditional validity theory. While challenges associated with the social context are acknowledged (Messick, 1996), experts debate whether these should be integrated into the validity criteria. This essay employs the terms "impact" and "consequences" in place of "consequential validity."

Tests wield influence on both macro (policy, societal, and institutional) and micro (individual) levels (Bachman & Palmer, 1996). Test washback impacts teaching and learning (Alderson & Wall, 1993). Particularly, high-stakes tests can shape teaching practices. Washback effects can be positive or negative (Alderson & Wall, 1993). When analyzing the social consequences of tests, McNamara and Roever (2006) propose exploring various dimensions of test use, including intended and actual consequences, as well as the underlying social and cultural values that inform test constructs and application.

Despite the significance of teachers' assessment literacy and their comprehension of the social and psychological implications of high-stake tests, uncertainty remains regarding the awareness of these consequences among EFL teachers, especially concerning nationally administered high-stake language tests. This study aims to assess the current level of Iranian EFL teachers' awareness of the social and psychological consequences of national high-stakes language tests and explore whether in-service training workshops influence their understanding of these test consequences. Consequently, the study's research questions are as follows:

1. What are the Iranian EFL teachers' cognitions of social and psychological consequences of high-stake English language tests administered at a national level?
2. Does test-consequences informed workshop affect the EFL teachers' cognitions of social and psychological consequences of high-stake English language tests?

Review of literature

Personal beliefs, attitudes, and cognitions of teachers can significantly impact teachers' classroom assessment practices (Borg, 1999, 2003). As defined by Fishbein and Ajzen (2010), personal beliefs, attitudes, and cognitions encompass the information individuals hold about objects, things, and other people around them. Lamont, et al. (2013) defined personal beliefs as propositional attitudes, which refer to individuals' attitudes towards propositions about objects, things, and other people. Dasgupta (2013) also argued that personal beliefs and cognitions could unintentionally influence individuals' decisions, judgments, and actions. Hence, teachers' perceptions, beliefs, and cognitions about assessment can significantly impact their implementation.

Numerous scholars have attempted to examine teachers' personal beliefs and cognitions regarding the main purposes of different assessment types (summative versus formative purposes). Teachers have demonstrated varied beliefs and cognitions about assessments in educational settings. For instance, Brown et al. (2011), using a quantitative research design, administered a questionnaire to investigate the beliefs and cognitions of 784 primary school teachers about assessment purposes in Queensland. Their findings revealed that teachers exhibited a stronger inclination towards formative purposes (improving learning) rather than summative purposes (grading and certification).

Similarly, Antoniou and James (2014), through document analyses, classroom observations, and semi-structured interviews, found that primary school teachers in Cyprus strongly believed that formative assessments played a pivotal role in enhancing effective teaching and learning. They also observed that teachers held different cognitions about the utility of classroom-developed assessments versus large-scale standardized tests, where large-scale testing typically serves summative functions while classroom-based assessment serves both formative and summative goals.

Leighton et al. (2010), employing a questionnaire to explore secondary school teachers' perceptions of assessments in Alberta, Canada, found that teachers believed their classroom assessment tasks yielded more diagnostic information (pertaining to the learning process, impact on meaningful learning, and use of learning strategies) than large-scale tests. This study concluded that teachers favored classroom-developed assessments over large-scale testing.

Gullickson (1984) surveyed 391 school teachers in rural Midwestern states in the USA regarding their cognitions and beliefs about the instructional use of tests. The findings indicated that teachers perceived traditional assessments, such as tests, as the best method to evaluate students' learning due to its perceived enhancement of instruction, increased student effort, influence on self-concepts, and encouragement of competition among students.

In contrast, Xu and Liu (2009) used an interview to explore an EFL college teacher's assessment cognition and reported that the teacher believed tests were the superior assessment method for measuring students' learning, as opposed to innovative assessment tasks. In contrast to the findings by Gullickson (1984) and Xu and Liu (2009), a questionnaire administered to school teachers revealed that teachers believed innovative assessments like construct-response tasks were the best format to assess students' learning, as they provided more comprehensive information about student achievements. Similarly, Schwager (1994), surveying school teachers about their assessment and instruction, found that teachers' cognitions and beliefs about assessments varied based on whether they viewed themselves as traditional or innovative teachers.

Cheng et al. (2004) found that ESL/EFL university teachers endorsed both innovative and traditional assessments. Canadian ESL/EFL university teachers employed innovative assessment methods more frequently than their Chinese and Hong Kong counterparts. Conversely, Chinese and Hong Kong teachers predominantly used traditional assessment methods like tests to evaluate students' learning.

Inbar-Lourie and Donitsa-Schmidt (2009) surveyed 113 EFL school teachers in Israel about the factors underlying the usage of innovative assessment and found that teachers' cognitions predicted the use of innovative assessment to evaluate student learning. The divergent findings among these studies regarding teachers' beliefs and cognitions about using traditional and innovative assessment methods were influenced by assessment purposes and institutional values and culture.

Teachers' differences in cognitions and beliefs about the utility of quality assurance procedures in assessment practices have also been observed. For instance, Gullickson (1984) and Gullickson and Ellwein (1985), using a questionnaire to examine school teachers' assessment cognitions, beliefs, and attitudes, reported that teachers viewed conducting statistical analyses of test scores (calculating reliability or item analyses) as impractical. In a subsequent study, Gullickson (1986) administered a questionnaire to 24 professors teaching an educational measurement course and 360 school teachers about their perspectives on pre-service educational measurement courses. The professors believed that statistical analyses of test scores were important, whereas teachers deemed such analyses unnecessary. Oescher and Kirby (1990) similarly reported that teachers believed their tests were reliable and valid, even without conducting statistical analyses of test scores.

In contrast, King (2010), through a questionnaire examining school teachers' and administrators' cognitions and beliefs about assessments, reported that teachers and administrators had more favorable attitudes towards educational statistics and considered conducting statistical analyses of test scores beneficial. Despite the potential benefits of Critical Assessment Literacy (CLA) in language education and assessment, there remains a scarcity of research examining the role of CLA in language assessment. Most studies have focused on the theoretical aspects of CLA, as evidenced by the works of Lynch (2001) and Shohamy (2001a). As of the current body of research, only two empirical investigations have sought to scrutinize the practical implications of CLA (Javidanmehr & Rashidi, 2011; Tahmasebi & Yamini, 2013).

The study conducted by Tahmasebi and Yamini (2013) aimed to examine the influence and significance of stakeholders, such as educators, learners, and guardians, in the formulation, implementation, and interpretation of the Iranian University Entrance Examinations (IUEE). The researchers employed Shohamy's (2001b) proposed CLA framework as the theoretical foundation for their study. The researchers collected data to address inquiries within three categories: factual, behavioral, and attitudinal. This was achieved using a Likert-scale survey instrument. The research supported Shohamy's perspective that assessments are potent instruments that advance the goals and agendas of influential entities.

The study by Javidanmehr and Rashidi (2011) aimed to implement Shohamy's proposed principles of CLA in a practical setting, with the goal of fostering a more egalitarian environment for language assessment. By selecting three principles of collaborative learning assessment (CLA), the researchers aimed to prioritize the incorporation of learners' perspectives into the assessment process. The study by Kiani et al. (2009) investigated the effects of administering English for specific purposes (ESP) tests as part of the admission process for postgraduate universities in Iran. The research aimed to examine the impact of these tests on both stakeholders and society. A qualitative research approach was used, involving 31 postgraduate students pursuing master's and doctoral degrees and five ESP instructors.

In summary, studies that examined teachers' personal beliefs and cognitions about assessment mainly utilized questionnaires with rating scales. Occasional use of classroom observations, document analyses, and interview techniques was also observed. Studies using rating scales provided limited insight into the developmental progression of teachers' personal beliefs and cognitions about assessment, as they often interpreted results based on midpoint scores. Therefore, further research is required to assess the levels of evolution of teachers' personal beliefs and cognitions about assessment.

Methodology

Participants

The study was conducted in two main phases, each involving a distinct group of participants. In the first phase, the focus was on examining the current state of social and psychological consequences associated with high-stake tests. For this purpose, 300 ELT teachers from various nationwide institutes were selected through a convenient sampling method to assess their perceptions of these consequences. The questionnaires were distributed, and a return rate of 70% was achieved, with 210 participants responding.

Moving on to the second phase, the objective was to evaluate the potential impact of in-service training workshops on non-native EFL teachers' perceptions of the social and psychological consequences of language tests. To achieve this, 39 EFL teachers with the lowest scores on the social and psychological consequences components were invited to participate in the workshop. Eventually, 30 of them (12 females and 18 males) agreed to join. The workshop took place on January 20, 2023, at Zagros Academic Institute in Tehran, Iran. Following the workshop, the participants were given a 2-week interval before they were invited once again to complete the instrument assessing their perceptions of the social and psychological consequences of language tests.

Instrumentation

In this study, the scale for social and psychological consequences of high-stake tests developed and validated by Alibakhshi et al. (2013) was used. It consists of four components: positive psychological consequences (4 items), negative psychological consequences (7 items), positive social consequences (4 items), and negative social consequences (4) items. The researchers developed the questionnaire through interviews with the test takers and reading the related studies and validated it using confirmatory factor analysis. The authors reported that the scale enjoyed an acceptable level of validity and reliability. In the piloting stage, the reliability of the scale was estimated using Cronbach's alpha, and the internal consistency was reported to be 0.87, which seems to be acceptable.

Treatment materials for the workshop

To design a teacher education course informed by the consequences of language tests, specific types of materials were needed in line with transformative teacher education. Therefore, the materials were developed or adapted based on the consequences of high-stake tests reported by Alibakhshi et al. (2013) and the related studies. The content of the materials for the in-service consequential validity-informed workshop was selected carefully to enhance teachers' awareness of the concept of consequential validity, the social consequences of language tests, and the psychological aspects of language tests to improve teachers' cognitions about the language test consequences. The workshop was held in two 90-min sessions, a total of 3 h. Each session generally followed a regular pattern of reflection, discussion, and action. Also, the participants were assigned readings on the covered materials prior to each session to encourage them to reflect, discuss, and act, accordingly.

Based on the developed syllabus for the workshop, the first session of the instruction was allocated to an introduction to types of validity, critical language testing, and assessment. The session aimed to provide the participants with valuable insights into the theoretical aspects of consequential validity and sharper their focus for subsequent sessions and further considerations. The second session dealt with the social and psychological consequences of language tests. Each consequence was defined, explained, and exemplified to the teacher.

Data collection procedure

The research was undertaken in two general phases, and each phase was completed through a number of steps. The first phase addressed investigating EFL teachers' cognitions of the social and psychological consequences of high-stake language tests. The instrument was administered to 3100 EFL teachers at different institutes and universities selected through a convenient sampling procedure. The returned questionnaires were collected, coded, and analyzed. Then, the teachers who obtained the lowest scores were screened and were invited to take part in the workshop. Those who were selected for the treatment phase were informed about the consequential validity, test consequences, and negative and positive consequences of the tests.

In the second phase, the researcher made attempts to investigate the impact of a consequences-informed workshop on 30 EFL teachers' cognitions. To achieve this, 30 EFL teachers were selected to take part in the study.

Then, the instructional sessions were held. The whole workshop took place in two sessions of 90 min (total of 3 h), and all sessions were specifically allocated to consequential validity conceptual and practical dimensions. Before the instructional phase actually began, the participants were asked to study some reading papers on critical language assessment so as to gain more insights into related issues and to critically analyze them through their own first-hand experiences. It could encourage them to incorporate the new readings from the literature at hand and their previous beliefs and experiences. It is thought that this prior reading of the instructional materials could help the researcher strike a trade-off between his role as a teacher educator and the participant teachers in terms of power distribution. As a result, a more cordial, cooperative, and democratic atmosphere with a more dialogic discourse structure was developed. At the end of the instructional session, teachers' cognition about test consequences was further explored by administering the instrument for test consequences.

Results

Results are presented in two sections: results for research question 1 and results for question 2.

Results for question 1

The first research question aimed at investigating the Iranian EFL teachers' cognitions of high-stake language tests' social and psychological consequences. Results are presented in Table 1.

As seen in Table 1, the mean scores of the participants on items 1, 2, 3, 4, and 5 which are related to the positive consequences and 12, which is related to negative consequences, exceeded the cutoff point which was set to be 2.5 (the hypothetical average). Results of one-sample t tests showed that the differences between the sample means and the hypothetical means of the population were statistically significant ($p=0.001$, df = 209) indicating that the EFL teachers are aware of the positive consequences of language high-stake tests. That is, they know that preparing for high-stake tests fosters the test-takers' learner autonomy ($M=3.6$, SD = , 1.11T = , $p=0.001$), good performance on language tests increases the test-takers' passion for learning ($M=3.7$, SD = 1.04,

Table 1 Teachers' cognitions of positive psychological consequences of high-stake language tests

Items	Statistics		*t* tests		
	Mean	SD	*T*	df	*P*
1. Preparing for high-stake tests fosters the test-takers' learner autonomy	3.60	1.11	4.1	209	0.001
2. Good performance on language tests increases the test takers' passion for learning	3.51	1.05	3.6	209	0.001
3. Achievement on high-stake tests increases the test takers' self-efficacy	3.70	1.04	3.4	209	0.001
4. Achievement on high-stake tests increases the test takers' self-confidence	3.40	1.01	3.3	209	0.001
5. Achievement on high-stake tests increases the test takers' positive academic emotions (hope, joy, etc.)	3.30	1.00	3.6	209	0.001
6. Test scores might negatively affect the test takers' mental health	2.5	1.16	0.78	209	0.63
7. Test scores might negatively affect the test takers' test anxiety	2.7	1.10	0.76	209	0.25
8. Test scores might negatively affect the test takers' foreign language anxiety	2.6	1.10	0.74	209	0.12
9. Test scores might negatively affect the test takers' mental health	2.43	1.13	0.83	209	0.16
10. Performance on language tests might lead to the test takers' academic burnout	2.32	1.09	0.56	209	0.32
11. Performance on language tests might increase the test takers' negative emotions	2.43	1.07	0.63	209	0.33
12. Performance on language tests might decrease the test takers' self-confidence	3.4	1.12	4.3	209	0.001

$T=3.3$, $p=0.001$), achievement on high-stake tests increases the test takers' self-efficacy ($M=3.5$, $\text{SD}=1.05$, $T=$, $p=0.001$), achievement on high-stake tests increases the test takers' self-confidence ($M=3.4$, $\text{SD}=1.01$, $T=3.6$, $p=0.001$), achievement on high-stake tests increases the test takers' positive academic ($M=3.3$, $\text{SD}=1.00$, $T=$, $p=0.001$), and performance on language tests might decrease the test takers' self-confidence ($M=3.4$, $\text{SD}=1.12$, $T=4.3$, $p=0.001$).

However, with regard to the negative consequences (items 6–11), it is seen that mean scores of the teachers fell either below or very close to the hypothetical mean of the population which was set be 2.5. Results of *t* tests ($p>0.05$) also verified that the EFL teachers are not well aware of the following negative psychological consequences:

A) Test scores might negatively affect the test takers' mental health ($M=2.5$, $\text{SD}=1.16$, $T=0.78$, $p=0.63$).
B) Test scores might negatively affect the test takers' test anxiety ($M=2.7$, $\text{SD}=1.1$, $T=0.76$, $p=0.25$).
C) Test scores might negatively affect the test takers' foreign language anxiety ($M=2.6$, $\text{SD}=1.1$, $T=0.74$, $p=0.12$).
D) Test scores might negatively affect the test takers' mental health ($M=2.6$, $\text{SD}=1.1$, $T=0.74$, $p=0.12$).
E) Performance on language tests might increase the test takers' negative emotions ($M=2.43$, $\text{SD}=1.7$, $T=0.63$, $p=033$)

As seen in Table 2, the mean scores of the sample on items 1, 2, 4, 6., 7, 11, 12, and 13 exceeded the hypothetical means of the population ($M=2.5$) and results of *t* tests showed that the differences between the sample and population means are statistically significant ($p=0.001$), favoring the sample. Therefore, it could be argued that EFL teachers know well about some of the social consequences of high-stake language tests, more

Table 2 Teachers' cognition of the social consequences of language high-stake tests

Items	**Statistics**		***t* tests**		
	Mean	**SD**	***T***	**df**	***P***
1. High-stake tests might lead to the test-takers' deprivation from higher education	3.60	1.11	4.1	209	0.001
2. High-stake tests might be biased toward one specific gender	3.42	1.03	3.6	209	0.001
3. Test items and tasks might be biased to test takers from specific ethnicities and cultural backgrounds	2.6	1.03	3.4	209	0.44
4. Test tasks and items might not be suitable for test takers with physical disabilities	3.40	1.01	4.32	209	0.001
5. Test items and tasks might lead to educational inequalities	2.56	1.06	0.56	209	0.36
6. Performance on high-stake tests affects the test takers' chances of migration to English-speaking countries	3.32	0.94	2.6	209	0.001
7. The test takers' scores on the high-stake tests affect the test takers' entry to famous universities	3.20	0.86	3.6	209	0.001
8. The test takers' scores affect their chances of finding a good job	2.53	1.03	0.59	209	0.33
9. Migrants with low band scores on high-stake tests might be marginalized	2.63	1.06	0.78	209	0.24
10. The test-takers' scores and performance on language tests affect their incomes	2.49	1.09	0.85	209	0.23
11. High-stake tests are good sources of English-speaking countries	3.3	0.98	3.32	209	0.001
12. Taking high-stake tests is very expensive for students from low-income countries	3.46	0.96	4.23	209	0.001
13. Test takers have to spend much time and money to get prepared for high-stake tests	3.39	1.1	2.96	209	0.001

specifically they know that high-stake tests can yield consequences that extend beyond mere assessment, potentially leading to the exclusion of test-takers from higher education opportunities. Moreover, these tests may exhibit biases favoring specific genders, thereby raising concerns about their fairness and equity. In addition, the suitability of test tasks and items might not adequately accommodate individuals with physical disabilities, accentuating the need for inclusive assessment practices.

The impact of high-stake test scores reverberates significantly, influencing test-takers' admissions into prestigious universities. Despite being regarded as valuable benchmarks in English-speaking countries, these tests carry the weight of several challenges. For instance, their accessibility comes at a cost, often rendering them financially burdensome for students hailing from low-income countries.

Preparation for high-stake tests demands substantial investment, both in terms of time and financial resources, from test-takers. As a result, the process of getting ready for these assessments can prove demanding. The cumulative effect of these challenges underscores the complex interplay between high-stake tests, educational opportunities, and socio-economic factors.

However, it is seen that the sample means on other items (3, 5, 8, 9, 10) fell either below the hypothetical means of the population ($M = 2.5$) or very close to it. Results of t tests also showed that the differences are not statistically significant ($p > 0.05$). Therefore, it can be strongly argued that EFL teachers are not well aware that the design of test items and tasks could potentially harbor biases that favor test-takers from certain ethnicities and cultural contexts. Consequently, this bias has the potential to perpetuate educational disparities among different groups. The implications extend further, as the scores obtained by test-takers influence their prospects of securing desirable employment opportunities. In the case of migrants, particularly those who attain low band scores on high-stake language tests, the consequences can be particularly stark. Such individuals might face marginalization, exacerbating challenges related to integration and social inclusion. Furthermore, the connection between test-taker performance on language assessments and their earning potential underscores the broader societal implications of language proficiency.

Research question 2

Research question 2 aimed at investigating whether test consequences-informed workshop course has a statistically significant effect on EFL language teachers' cognition about social and psychological consequences of language tests. To do so, the researcher compared the scores of the participants on the scales administered before and after the treatment. Results of the descriptive and inferential statistics are presented in the following Table 3.

As shown in the Table 3, the mean scores of the teachers on the total and individual items of the psychological consequences of the high-stake tests after the treatment exceeded their mean scores on the total and individual items of the instrument before the treatment. Results of inferential statistics (paired *t* test) also verified that the difference between the mean scores on the total component and individual items is statistically significant ($p = 0.001$). Therefore, it could be inferred that teacher education courses significantly increased the teachers' knowledge of the psychological consequences of language tests. The results of the statistical procedures for the *social* component of the scale are presented in Table 4.

As shown in Table 4, the mean scores of the teachers on the total and individual items of the social consequences of the high-stake tests after the treatment exceeded their

Table 3 T test for teacher's knowledge of psychological consequences of high-stakes English language tests before and after the treatment

Item	Statement	M1	M2	*T*	*P*
1	Preparing for high-stake tests fosters the test-takers' learning autonomy	2.2	3.8	7.2	0.001
2	Good performance on language tests increases the test takers' passion for learning	2.8	3.9	3.1	0.001
3	Achievement on high-stake tests increases the test takers' self-efficacy	2.2	4.5	13.3	0.001
4	Achievement on high-stake tests increases the test takers' self-confidence	2.3	3.8	8.3	0.001
5	Achievement on high-stakes tests increases the test takers' positive academic emotions (hope, joy, etc.)	2.2	3.9	7.4	0.001
6	Test scores might negatively affect the test takers' mental health	2.1	3.4	5.5	0.001
7	Test scores might negatively affect the test takers' test anxiety	2.1	3.8	10.3	0.001
8	Test scores might negatively affect the test takers' foreign language anxiety	1.7	3.2	8.4	0.001
9	Test scores might negatively affect the test takers' mental health	2.2	3.8	8.41	0.001
10	Performance on language tests might lead to the test takers' academic burnout	2.2	3.8	8.4	0.001
11	Performance on language tests might increase the test takers' negative emotions	2.6	3.9	6.3	0.001
12	Performance on language tests might decrease the test takers' self-confidence	2.1	3.8	10.5	0.001
	Total	27.7	47.3	17.8	0.001

Table 4 *T* test for teacher's knowledge of social consequences of high-stake English language tests before and after the treatment

Item	Statement	M1	M2	*T*	*P*
1	High-stake tests might lead to the test-takers deprivation from higher education	2.8	3.6	4.1	0.001
2	High-stake tests might be biased towards one specific gender	2.5	4.2	3.5	0.001
3	Test items and tasks might be biased to test takers from specific ethnicities and cultural backgrounds	2.4	3.9	3.6	0.001
4	Test tasks and items might not be suitable for test takers with physical disabilities	2.7	3.8	2.9	0.001
5	Test items and tasks might lead to educational inequalities	2.6	4	7.6	0.001
6	Performance on high-stake tests affects the test takers' chances of migration to English-speaking countries	2.4	3.8	6.3	0.001
7	The test takers' scores on the high-stakes tests affect the test takers' entry to famous universities	2.1	3.7	6.6	0.001
8	The test takers' scores affect their chances of finding a good job	2.2	3.8	6.5	0.001
9	Migrants with low band scores on high-stake tests might be marginalized	2.3	3.8	8.41	0.001
10	The test-takers' scores and performance on language tests affect their incomes	2.4	3.8	8.4	0.001
11	High-stake tests are good sources of English-speaking countries	2.7	3.9	6.3	0.001
12	Taking high-stake tests is very expensive for students from low-income countries	2.5	3.8	10.5	0.001
13	Test takers have to spend much time and money to get prepared for high-stake tests	2.5	3.9	11.2	0.001
	Total	32.1	50	17.8	0.001

mean scores on the total and individual items of the instrument before the treatment. Results of inferential statistics (paired *t* test) also confirmed that the difference between the mean scores on the total component and individual items is statistically significant ($p = 0.001$). Therefore, it could be inferred that teacher education courses significantly increased the teachers' knowledge of the social consequences of the language tests.

Discussion

The first objective of the study was to investigate the Iranian EFL teachers' cognition of the high-stake tests. The findings of the study showed that Iranian EFL teachers are well aware of the positive psychological consequences of English language tests, but they are not well aware of some of the negative psychological consequences. The findings are consistent with Spolsky (1981) who raised awareness of the direct consequences of tests/assessments by posing two basic questions, both of which spurred debate in discussions surrounding assessment and which ultimately led to practice taking place in the field: "How sure are you of your decision?" and "How sure are you of the evidence that you're using to make that decision?" (p. 19). Messick's (1989) "unitary validity model" was the first attempt to propose an assessment model which forged a connection between assessment and its impact via introducing the term *consequential validity*. He points out:

> *The consequential aspect of construct validity includes evidence and rationale for evaluating the intended and unintended consequences of score interpretation and use in both the short and long term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning (1996, p. 251).*

This finding is also in line with Bachman and Palmer's (1996) concept of *impact* in their "usefulness" framework. It also echoes *the impact-consequential validity* and *social* *consequence theory* introduced by Lynch (2001) as a basic quality of his “test fairness” framework. However, assessment literacy is entirely about assessment consequences referring to the linkage of assessment to psychological, social, and political variables which may have bearings on such concepts and areas as curriculum design, ethicality, social classes, bureaucracy, politics, and knowledge (Shohamy, 2017). Critical assessment literacy treats assessment practice, not as a neutral activity but as something completely social. That is, assessment tools are “connected and embedded in political, social, and educational contexts.” (Shohamy, 2007, p. 117). According to the fundamental tenets of CLA, the quality of tests and assessment “is not judged merely by their psychometric traits but rather about to their impact, ethicality, fairness, values, and consequences.” (Shohamy, 2007, p. 117).

The results are also consistent with the results of studies by some scholars (Tavassoli & Farhadi, 2018; Inbar-Lourie, 2008; Malone, 2013; Taylor, 2013; Tajeddin, et al., 2022) who have reported dissatisfaction with the status of EFL teachers’ language assessment Knowledge. A look at the research published in the specialist literature shows that most teachers have indicated that they do not feel comfortable using measurement and assessment principles in practice. Some scholars even claim that teachers evaluate the performance of students even when they have little or no professional training (Bachman, 2014). Insufficient assessment knowledge by teachers is often attributed to a failure to receive effective assessment training during their preparatory or in-service teacher training programs.

The second research question delved into the effects of a test consequence-informed workshop on shaping English language teachers’ cognition about the social and psychological consequences of language tests. The pertinent results of both descriptive and inferential statistics indicated that their cognition about social and psychological consequences on all total components and individual items significantly improved in post-test administration of the questionnaire, which was in line with the findings of the previous studies on teachers’ critical consciousness, cognitions, and practices (Borg, 2003, 2011; Khatib & Miri, 2016; Miri et al., 2017b; Tajeddin, et al., 2022).

Borg (2011) carried out a qualitative longitudinal study to investigate the effect of an in-service teacher education course on teachers’ beliefs about teaching pedagogy. The findings showed that changes have been made to teachers’ prior beliefs about language teaching and learning; however, “despite this evidence of impact, the data also suggest that the in-service course studied here could have engaged teachers in a more productive and sustained examination of their beliefs” (p. 370). Similarly, Busch (2010) carried out a study to examine the impact of an introductory course on issues relevant to second language acquisition on pre-service teachers’ beliefs. The findings reflected significant changes in teachers’ beliefs regarding some key aspects of SLA such as error correction, culture and language, and the complexity of language learning. Khatib and Miri (2016) conducted a study on the effect of a CP-informed teacher education program on teachers’ talk to foster multimodality in EFL classrooms. They audio- and video-recorded two teaching sessions of the participating teachers prior and after the course to investigate the possible changes in teachers moves to curb or improve multimodality. They concluded that the course significantly improved the participating teachers’ moves to enhance the multimodality after the course. In addition, Miri et al. (2017a) explored the role of a teacher education course informed by the principles of critical pedagogy on teachers' cognitions and practices on using L1 in EFL classrooms. Interviews and stimulated recall sessions were used to examine the changes in the participating teachers' cognition and practice and the findings revealed that the course was highly effective in (re) shaping teachers' cognitions about L1 use and in using L1 in their own classrooms.

The changes in teachers' perceptions of the social and psychological consequences of language testing can be attributed to a variety of reasons. As a good example of this, the findings could be partially explained by Zones of Proximal Teacher Development (ZPTD) (Warford, 2011), originally borrowed from Vygotsky's concept of one of proximal development (ZPD) and relating to the gap and distance between learners' and actual ones' developmental level, which is what a given individual can achieve alone, and potential developmental level, which refers to what the learner can do and achieve when supported by an adult or a more able peer (Warford, 2011). It seems that a test consequence-informed workshop could help EFL teachers develop an insight into different aspects of test consequences as prompts and hints were given to the participating teachers in a reflective dialogue context. That is, the teachers pooled their cognitive resources to collectively develop a critical awareness of language assessment that might not have emerged without collaborative dialogues.

Conclusions and implications

Based on the findings of the study and in line with Shohamy (1997), it can be concluded that test consequences allude to possible washback, both positive which is generally intended and negative which is normally unintended occur in the educational context; for example, tests can govern textbook as well as a curriculum as educational devices (Shohamy, 1999). It can also be concluded that teachers need to know how a language test might lead to negative social and psychological consequences. Teachers, if not well aware of the unintended negative consequences of language tests, cannot teach in line with the principles of critical pedagogy and reflective teaching. When tests are designed with awareness and understanding of some factors such as the learning contexts, students, and the contents, positive washback is more likely to appear (Xerri & Vella Briffa, 2018). Apart from the educational consequences, if it is acknowledged that L2 learning is chiefly a social psychological event, it is merely natural that social-psychological variables should be given central attention(Shohamy, 1999).

Although Iranian EFL teachers are aware of the social, psychological, and educational consequences of language tests, they need to develop their cognition of test use consequences, test fairness, test policies, and governments' ideological and cultural trends in test development, use, and interpretations. Teachers' cognition about critical language assessment and test consequences can be shaped through explicit instructions in pre-service and in-service training courses. It can also be concluded insufficient cognition of test consequences affects the test takers' performance and will certainly lead to negative consequences including test bias and test anxiety. Therefore, the inclusion of pre-service and in-service training courses on critical language assessment literacy can both increase cognition of test consequence status among EFL teachers and keep them updated.

The findings of this study have profound implications for teacher training and professional development in the context of Iranian EFL education. The research reveals that Iranian EFL teachers demonstrate an awareness of the positive psychological effects of high-stake English language tests, but they appear to lack a comprehensive understanding of the potential negative psychological consequences. This underscores the urgent need for targeted initiatives aimed at enhancing teachers' assessment literacy. Assessment literacy must go beyond a mere understanding of psychometric traits and encompass a broader comprehension of the implications and outcomes of assessment practices. By equipping teachers with a more holistic understanding of assessment consequences, institutions can ensure that educators make informed decisions regarding test administration, interpretation, and their broader implications. Professional development programs can play a pivotal role in addressing these gaps in assessment literacy.

The positive outcomes observed in the study due to the test consequences-informed workshop highlight the efficacy of such interventions. These workshops offer a focused platform for enhancing teachers' perceptions of the social and psychological consequences of language tests. Thus, institutions should consider integrating similar workshops as a regular component of their professional development offerings. This would enable educators to continuously update their assessment knowledge, aligning their practices with the evolving understanding of assessment's broader impact.

Furthermore, the study emphasizes the power of collaborative learning and reflection. The workshop facilitated collaborative dialogues among teachers, fostering critical awareness of language assessment. Educational institutions can harness this approach by encouraging platforms for teachers to engage in discussions and share insights on assessment practices. Such collaborative learning environments can help educators collectively enhance their assessment literacy and better understand the contextual factors that influence assessment practices.

However, the study also points out that the impact of professional development is not limited to a one-time event. Instead, sustained engagement with assessment literacy topics could yield more substantial and lasting changes. Therefore, institutions should adopt a long-term perspective on professional development, viewing it as an ongoing process to continuously improve teacher practices and beliefs.

Abbreviations

CFA	Confirmatory factor analysis
CLA	Critical assessment literacy
CLA	Collaborative learning assessment
EFA	Exploratory factor analysis
EFL	English as a foreign language
EPT	English proficiency test
ESP	English for specific purposes
L2	Second language
IUEE	Iranian University Entrance Examinations
ZPD	Zone of proximal development
ZPTD	Zones of Proximal Teacher Development

Acknowledgements

The authors would like to thank the participants of the study as well as the administrators of the language institutes.

Authors' contributions

All authors equally contributed to the study.

Funding

Funding is not applicable to this study.

Availability of data and materials
The data would be available upon the editors' request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 1 June 2023 Accepted: 3 October 2023
Published online: 28 November 2023

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Alibakhshi, G., & Rezaei, Mezajin S. (2013). On the consequences of the violation of critical pedagogy principles from Iranian EFL teacher trainers' perspectives. *Iranian Journal of Applied Language Studies, 5*(2), 1–28. https://doi.org/10.22111/IJALS.2015.1875

Alibakhshi, G., & Shahrakipour, H. (2014). The effect of self-assessment on EFL learners' receptive skills. *Journal Pendidikan Malaysia, 39*(1), 9–17.

Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability, 10*(4), 1–24.

Bachman, L. F. (2014). Ongoing challenges in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3, pp. 1586–1603). Oxford: Wiley.

Bachman, L., & Palmer, J. (1996). *Language testing in practice*. Oxford University Press.

Board, B. (2003). The power of tests: A critical perspective on the uses of language texts, Review of the article: *The power of tests: A critical perspective on the uses of language texts*, by Elena Shohamy (Longman, 2001). *Journal of Writing Assessment, 3*(1), 55–60.

Borg, S. (1999). Studying teacher cognition in second language grammar teaching. *System, 27*(1), 19–31.

Borg, S. (2003). Teacher cognition in language education: A review of research on what language teacher think, know, believe, and do. *Language Teaching, 36*(2), 81–109.

Borg, S. (2011). The impact of in-service teacher education on language teachers' beliefs. *System, 39*(3), 370–380.

Brown, G. T. L. (2019). Is the assessment for learning really an assessment? *Frontiers in Education, 4*(64), 1–7.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Longman.

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210–220.

Busch, D. (2010). Pre-service teacher beliefs about language learning: The second language acquisition course as an agent for change. *Language Teaching Research 14*(3), 318–337.

Butler, Y. G., Peng, X., & Lee, J. (2021). Young learners' voices: Towards a learner-centered approach to understanding language assessment literacy. *Language Testing, 38*(3), 429–455. https://doi.org/10.1177/0265532221992274

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/ EFL instructors‟ classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21*(3), 360–389.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (pp. 233–279). Elsevier.

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Corwin.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action Approach*. Psychology Press.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. *The Journal of Educational Research, 77*(4), 244–248.

Gullickson, A. R., & Ellwein, M. C. (1985). *Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice*. Educational Measurement: Issues and Practice.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational measurement, 4*(23), 354–347.

Inbar-Lourie, O. (2008). Language assessment culture. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (2nd ed., Vol. 7, pp. 285–299). New York: Springer.

Inbar-Lourie, O., & Donitsa-Schmidt, S. (2009). Exploring classroom assessment practices: The case of teachers of English as a foreign language. *Assessment in Education: Principles, Policy & Practice, 16*(2), 185–204.

Javidanmehr, Z., & Rashidi, N. (2011). Critical language assessment: Students' voices at the heart of educational system. *Elixir Psychology, 37*, 3740–3746.

Khatib, M., & Miri, M. (2016). Cultivating multivocality in language classrooms: Contribution of critical pedagogy-informed teacher education. *Critical Inquiry in Language Studies, 13*(2), 98–131.

Kiani, G. R., Alibakhshi, G., & Akbari, R. (2009). *On the consequential validity of ESP tests: a qualitative study in Iran*.

King, J. D. (2010). *Criterion-referenced assessment literacy of educators*. Unpublished doctoral thesis. The University of Southern Mississippi.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly, 17*(1), 100–120. https://doi.org/10.1080/15434303.2019.1674855

Lamont, S., Jeon, Y. H., & Chiarella, M. (2013). Health-care professionals' knowledge, attitudes and behaviours relating to patient capacity to consent to treatment: an integrative review. *Nursing ethics, 20*(6), 684–707.

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2020). Formative assessment as practice: The role of students' motivation. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2020.1765228

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom-versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice, 17*(1), 7–21.

Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly, 17*(2), 161–182. https://doi.org/10.1080/15434303.2019.1692347

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

Malone, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

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

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9–20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). MacMillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.

Miri, M., Alibakhshi, G., Kushki, A., & Bavarsad, P. S. (2017a). Going beyond one-to-one mediation in zone of proximal development (ZPD): concurrent and cumulative group dynamic assessment. *Eurasian Journal of Applied Linguistics, 3*(1), 1–24.

Miri, M., Alibakhshi, G., & Mostafaei-Alaei, M. (2017b). Reshaping teacher cognition about L1 use through critical ELT teacher education. *Critical Inquiry in Language Studies, 14*(1), 58–98.

Oescher, J., & Kirby, P. C. (1990). *Assessing Teacher-Made Tests in Secondary Math and Science Classrooms.*

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265–273.

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Pearson Education.

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7th ed.). McGraw-Hill.

Schwager, M. T. (1994). *Paradigms of assessment: What is authentic assessment in your world view?* University of California, Riverside.

Shohamy, E. (1997). Critical language testing and beyond. Plenary talk presented at the American Association of Applied Linguistics (AAAL) Meeting, Orlando, Florida. *Studies in Educational Evaluation, 24*(1), 331–45.

Shohamy, E. (1999). *Critical language testing, responsibilities of testers and rights of test takers*. Paper presented at the AERA Convention, Montreal.

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–392.

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

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

Shohamy, E. (2009). Language teachers as partners in crafting educational language policies? *Íkala, Revista De Lenguaje y Cultura, 14*(22), 45–67.

Shohamy, E. (2017). ELF and critical language testing. In J. Jenkins, M. Dewey, & W. Baker (Eds.), *The Routledge Handbook of English as a lingua franca* (pp. 583–593). Routledge.

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–30). Frankfurt am Main: Verlag Peter D. Lang.

Sultana N. (2019). Language assessment literacy: an uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, 9, Article 1. https://doi.org/10.1186/s40468-019-0077-8

Tahmasebi, S., & Yamini, M. (2013). *Power relations among different test parties from the perspective of critical language assessment*.

Tajeddin, Z., Khatib, M., & Mahdavi, M. (2022). Critical language assessment literacy of EFL teachers: Scale construction and validation. *Language Testing, 39*(4), 649–678.

Tavassoli, K., & Farhady, H. (2018). Assessment knowledge needs of EFL teachers. *Teaching English Language, 12*(2), 45–65.

Taylor, L. (2013). Communicating the theory, practice, and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41–63. https://doi.org/10.1080/15434303.2014.960046

Vogt, K., Tsagari, D., & Spanoudis, G. (2020). What do teachers think they want? A comparative study of inservice language teachers' beliefs on LAL training needs. *Language Assessment Quarterly, 17*(4), 386–409. https://doi.org/10.1080/15434303.2020.1781128

Warford, M. K. (2011). The zone of proximal teacher development. *Teaching and Teacher Education, 27*(2), 252–258.

Xerri, D., & Briffa, P. V. (Eds.). (2018). *Teacher involvement in high-stakes language testing*. Springer.

Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *Tesol Quarterly, 43*(3), 492–513.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.