RESEARCH

Open Access

What ethical requirements should be considered in language classroom assessment? insights from high school students

Zeinab Azizi[1] and Sayed M. Ismail[2*]

*Correspondence:
a.ismail@psau.edu.sa

[1] Teaching English and Linguistics Department, University of Ayatollah Ozma Borujerdi, Borujerd City, Iran
[2] College of Humanities and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

Abstract

There has been a growing call for listening to test-takers' voices across diverse assessment contexts. Within classroom assessment (CA), however, test-takers' voices on ethics are under-researched in the high school context of Iran. Hence, this study purported to disclose Iranian high school test-takers' ($n = 15$) perceptions of the ethical requirements in CA. For this purpose, a systematic thematic coding approach (constant-comparative method) was used to analyze the participants' perceptions. Findings yielded two overarching categories, including *do no harm* (e.g., establishing a supervision group, considering test-takers' individual differences, keeping test results confidential, and turning back test sheets with feedback) and *avoid score pollution* (e.g., using additional knowledge sources, using alternative assessment methods, clarifying grading criteria, avoiding unfamiliar contents and surprise items). The findings refer to a local gloss on global principles of ethics, which is hoped to map out specific dimensions of this important notion for diverse assessment contexts (e.g., high-stakes language testing and CA) and stakeholder groups (e.g., high school teachers, assessment developers, and education officials).

Keywords: Ethical requirements, Classroom assessment, A thematic coding analysis, High school test-takers' voices

Introduction

There has been a growing body of research in recent decades examining the ethical and social implications of language assessment for various stakeholder groups (Murchan & Siddiq, 2021). Ethical violations may occur, for instance, when test-takers are not adequately prepared for a novel or complex method of assessment (Gitsaki & Robby, 2018) or when they are assigned to a group that is too advanced for their competence (Brown and Harris, 2016). Ethics is "a set of accepted beliefs and practices meant to restrain behavior and promote the common good," according to one definition of the term (Tylor, 2013, p. 1). The principles of ethics are expressed in professional codes of ethics, such as the draft Code of Practice (ILTA, 2005), the International Language Testing Association's Code of Ethics (ILTA, 2000), and the Code of Practice (ALTA, 2005) and supporting quality assurance framework from the Association of Language Testers in Europe.

Other examples of professional codes of ethics include the Code of Ethics of the Association of Language Testers in Europe (ALTE, 2001). “Ethical behavior” (acting based on one’s judgment of an obligation) can be implemented by either practicing the principles of ethics (Davies, 1997) or by the building of an “ethical milieu” (Homan, 1991) “through a professional association (such as ILTA)” (Davies, 1997; Homan, 1991). “Ethical behavior” refers to the act of acting based on one’s judgment of commitment (Davies, 2008, p. 436).

The significance of ‘ethical behavior’ in the evaluation of test-takers has been widely acknowledged in assessment practices, and researchers have placed an emphasis on this value (e.g., Azizi, 2022; Green et al., 2007; Pope, 2006a, b; Popham, 2000). The use of ethical assessment practices has been shown by empirical research to effectively boost test-takers’ learning and performance in a variety of academic areas (Green & Johnson, 2010). Because of this crucial understanding, more and more professionals are becoming aware of the need to pay particular attention to the ethicality dimension (Mathew, 2004), which they do by theoretically addressing the significance of this key concept (e.g., Airasian, 2005; Davies, 2008; Hamp-Lyons, 2000; Rezai et al., 2022; Taylor & Nolen, 2005).

Recent years have seen a growth in the number of empirical studies that have taken a careful look at the concept of student evaluation in relation to the problem of ethics in high-stakes testing (Mathew, 2004; O’Loughlin, 2011). It is widely known that classroom assessment (CA)—that is, formative and summative assessment practices that teachers employ to evaluate student learning—has the ability to cultivate favorable perceptions and to improve student learning (Brown and Harris, 2016). As a result, it is absolutely ethical for research projects to investigate the nature of this significant setting from the perspective of ethics. In response to this request for inquiry, a few investigators have focused their attention on the ethical ethics that pertain to CA (e.g., Fan et al., 2017, 2020; Popham, 2000; Tierney, 2014). Some research produced empirical evidence of potential ethical difficulties across diverse testing stakeholder groups, such as pre-service teachers (Fan et al., 2019; Liu et al., 2016), university lecturers (Fan et al., 2017), in-service teachers (Pope, 2006a, b), and education stakeholders (Johnson et al., 2008). Research on test takers’ (students’) opinions and experiences of ethical behavior is, to this day, lacking in scope and depth. To put it another way, the voices of test-takers are conspicuously lacking from the assessment literature (Brown et al., 2020). In example, evidence of validity coming from the perspectives of the people who took the test has been judged to be irrelevant (Cheng, 2008; Cohen, 2006; Hamp-Lyons, 2000).

Despite the fact that these various studies have provided valuable insights for both relevant research and CA practices, it is clear that the authors have simply followed planned and predefined scenarios of ethics in the form of surveys inside CA (e.g., Fan et al., 2017; Green et al., 2007; Pope, 2006a, b). In addition, as Rasooli et al. (2018) point out, the use of only a quantitative paradigm to investigate the fairness of CA practices results in the omission of a wide variety of crucial CA procedures and practices. Although Fan et al. (2020) analyzed the perspectives of Chinese college test-takers on the ethicality of CA practices in a more recent study, they only used a couple of scenario-based items to judge the ethicality or unethicalness of some assessment practices. Although this is a more recent study, it was still conducted in China; exploring the critical viewpoints of test takers through the use of a qualitative study, which might provide rich data that was lacking from their analysis. The issue at hand is the fact that, as of right now, we only have a hazy understanding of the degree to which a group of stakeholders that is frequently ignored, test-takers themselves, can provide us with valuable insights about the principles of ethics in CA and how these perspectives might be different from the ones that have been reported so far for teachers. This is the root of the problem. The current study may expand our understanding by carefully examining test-takers' ethical viewpoints on the current practices of CA at the high school contexts of Iran, given that CA has gained substantial momentum in today's standards-based system of education (Harris and Brown, 2016) and paired with the growing call for listening to test-takers' voices (Cheng & DeLuca, 2011; Hamid, 2014; Rezai, 2022). Thus, the present study may contribute to the literature by further our understanding of the test-takers' perceptions of ethical requirements in the high school contexts of Iran.

Ethics in classroom assessment

Evaluation of test-takers' progress through the use of CA is an important component of the educational process. The CA, which can be completed in either a summative or formative manner, has two fundamental purposes: assessment of learning and assessment for learning (Doosti & Ahmadi Safa, 2021; Fan et al. 2020a; Green & Johnson, 2010). The Joint Committee on Standards for Educational Evaluation (JCSEE, 2003, 2015) is a group of educational researchers and scholars who have produced and presented Classroom Assessment Standards for teachers. This was done with the intention of achieving such valuable reasons. Cultural and linguistic variety, exceptionality and special education, objective and fair assessment, reliability and validity, and introspection all play a role in these standards that are grounded in research-based principles and theoretically-informed guidelines (JCSEE, 2015). The passage of time has resulted in other standards being added to the list, such as "avoiding score pollution," which refers to instances in which test results are inflated and do not adequately reflect the learning of test-takers (Fan et al., 2020). Rasooli et al. (2018) identified two significant facets that should be prioritized while attempting to minimize score pollution. "First and foremost, elements that are irrelevant to the construct being measured should not be included in test-takers' marks. Secondly, factors that underrepresent the construct being measured should not be included in test-takers' grades" (p. 171). Pope et al. (2009) emphasize the need of adhering to the concept of preventing score pollution in addition to the various other standards and principles in order to guarantee that the ethical requirements of the CA are satisfied.

In the context of the classroom, the goal of ethics is to guarantee that the assessment practices are both effective and fair. Some categories, such as communication regarding grading, confidentiality, grading practice, techniques of assessment, test administration, and standardized test preparation, should be taken into consideration and put into practice by teachers and test-makers in order to ensure the ethical nature of CA (Fan et al., 2020; Green et al., 2007). "Student assessments should be ethical, fair, practicable, and accurate, "according to the fair and unbiased assessment, which is one of the basic five quality standards (JCSEE, 2015, p. 3). Recent research has shown that this aspect is receiving a significant amount of attention (e.g., Oosterhof, 2009; Popham, 2017; Rasooli et al., 2018; Waugh & Gronlund, 2013). For instance, Waugh and Gronlund (2013) discovered that "personal bias and the halo effect" (p. 169) are two key challenges to the process of providing a fair assessment. Also, Rasooli et al. (2018) come to the conclusion that student evaluation should be designed and administered in light of the factors (such as student ability, effort, attendance, and attitude) that maximize the test accuracy in measuring student learning. These factors include student ability, effort, attendance, and attitude.

"Communication about test processes" is a part of CA ethics that is considered to be of the utmost importance. As its name suggests, the activity known as "communication about test processes" requires teachers to communicate with test-takers about the substance of tests, testing procedures, test administrations, grading criteria, and the interpretation of test results (Airasian, 2005; Ory & Ryan, 1993; Stiggins et al., 1989). To put it another way, test-takers need to be aware of the steps that are taken from defining the construct, sampling language behavior, measuring and reporting test performance so that decisions may be made based on test scores. "Confidentiality" is another element of the CA code of ethics that is absolutely necessary. If the results of an assessment practice are made public, it could potentially undermine the assessment's ability to give an accurate representation of the test takers' performance. When it comes to ensuring test-takers' privacy and the secrecy of their test scores, educators are tasked with the responsibility of maintaining confidentiality. Any individual who does not have the right to be informed about test findings should not be made aware of them nor should they be publicized publicly (JCSEE, 2015). The outcomes of a study that was carried out by Brookhart and Nitko (2008) provided proof that professors at universities ought to protect the privacy of their test-takers and maintain the confidentiality of their exam scores. Another factor that adds to CA's reputation for ethical ethicality is its "grading practice." The term "grading practice" refers to telling test-takers about the weight of different elements of the learning materials for the final evaluation, providing clarity regarding the grading rubric, and returning test papers to test-takers (JCSEE, 2015). When grading criteria are more accurate and fair, they bring about a greater number of benefits for various testing parties. This can be traced back to the fact that the test-takers' grades serve as a substantial source of motivation for them to continue their education (Brookhart & Nitko, 2008). In addition, the practices of grading give test-takers with feedback on their learning, which clarifies what they have comprehended, what they have not grasped, and where there is room for improvement. In addition, grading practices provide feedback to teachers on the learning of pupils, information that can be used to shape future decisions on instruction (Ory & Ryan, 1993). "Administering a variety of assessment methodologies" is the second component of assessment practices that should adhere to ethical standards. Since it is obvious that a single assessment strategy is unable to quantify student learning in an accurate manner, the utilization of a variety of assessment practices is absolutely necessary (Brookhart & Nitko, 2008; Gronlund, 2003). The fundamental reason for this is due to the fact that various methods of assessment each have their own set of benefits and drawbacks (Backman and Palmer, 2010; Green & Johnson, 2010). For example, Green and Johnson (2010) discovered that even though multiple-choice tests have a high level of feasibility and can cover a broad range of subject matter areas, they are not very useful for measuring test-takers' higher-level cognitive skills, such as creativity. This was the case despite the fact that these tests enjoy a high level of practicability. The 'administration of tests' is the next component of conducting ethically sound evaluation practices (Brookhart & Nitko, 2008). The administration of the test encompasses a variety of aspects, such as the arrangement of the physical setting (for example, light, temperature, ambient noise, ventilation, and minimal distractions) and the behaviors of the test administrators (for example, proctors should not unduly intervene in the performance of the test-takers) (Popham, 1991). In this regard, Sax (1974) suggests that teachers try to standardize all components of exams so that test-takers' performance is not adversely affected by extraneous conceptions. This is done to prevent test-takers from answering questions incorrectly. "Standardized test preparation" is the final aspect that contributes to the ethicality of assessment practices (Fan et al. 2020a). The pupils' preparation for exams should be considered of the utmost priority. Because when test-takers are aware of the specifics of the assessment practices, they are able to better prepare themselves and more effectively demonstrate their abilities.

Research on test-takers' perceptions of tests

The need to hear test-takers' voices has increased during the past 10 years. (e.g., Cheng & DeLuca, 2011; Fan et al. 2020a; Hamid, 2014a; Hamid et al., 2019; Murrillo and Hidalgo, 2017; Pearson, 2019). The belief that test-takers and the viewpoints they bring to the table should be incorporated into assessment practices is driving this increased interest. It is imperative that test-takers not only be regarded as subjects for the purpose of the examination, but also as individuals who are participating in language assessments (Hamid & Hoang, 2018). Moreover, "extensive discussion of the test-takers' roles in test development and evaluation, the rising number of empirical research that seek to integrate test-takers' voices into testing procedures, and the numerous validation frameworks that have advocated for including test-taker viewpoints" all contribute to this heightened focus on test-takers' perceptions and experiences (Hamid & Hoang, 2018, p. 1).

A point that is even more significant is that the new models of test validation have allotted a specific place for the voices of test takers. As an illustration, the point of departure of assessment practices is considered to be the test-takers in the socio-cognitive paradigm that was developed by Weir (2005). It emphasizes the importance of taking into account the physical, psychological, and experiential aspects of those who are taking the test (Cheng & DeLuca, 2011). According to Bachman and Palmer (2010), the viewpoints of test takers can be used to document the validity of a test. The viewpoints of test takers can be extremely useful when creating tests, improving the judgments that are made based on test scores, and boosting the beneficial repercussions of test scores for a variety of stakeholders, most notably test takers. In a similar vein, within the critical perspective of language testing, it has been proposed that in order to make the exam more people-oriented, one of the key knowledge sources for test-makers should be the perceptions of the people who are taking the test (Fan et al. 2020a; Shohamy, 2001b; 2007, 2013). It is stated that the democratic and humane principles of critical language testing cannot be put into effect unless sufficient attention is paid to the perceptions of those who take the tests.

Both theoretical discussions (e.g., Hall, 2009; Hamid, 2016; Hamid & Hoang, 2018; Pearson, 2019; Uysal, 2009) and empirical studies have paid a significant amount of attention in recent years to the critical voices of test-takers regarding high-stakes and low-stakes assessment practices (e.g., Ahmadi, 2021; Cheng DeLuca, 2011; Fan et al. 2020a; Hamid, 2016; Hamid et al., 2019; Hyatt, 2013; Murrillo and Hidalgo, 2017; Xiao & Carless, 2013). For instance, Cheng and DeLuca (2011) conducted an investigation on the viewpoints of 59 test-takers when developing large-scale English language exams. This was one of the earliest attempts to take into account the test-takers' conceptions regarding the validity of the tests. Their findings suggested that the test-takers' voices could disclose essential features of language assessment, which has valuable implications for test-developers, test-administrators, and test-users. In a further piece of research that was conducted out by Hamid (2016), the test-takers' perceptions regarding the use of the IELTS exam as a "policy tool" for making decisions concerning the test-takers' life were looked into. His findings demonstrated that the test-takers' experiences and perceptions can legitimately call into question the reliability claims made by the testing agency by drawing attention to the commercial motivations at the core of their policy, in addition to raising theoretical, professional, and ethical questions. More recently, Pearson (2019) dug into its critical perspectives on the development and administration from test-takers' perspectives. In addition to acknowledging the fact that the IELTS test is a high-stakes, high-pressure test that acts as a global gatekeeping test that regulates the migration and academic study of people across the world, Pearson (2019) dug into its critical perspectives on the development and administration from test-takers' perspectives. According to the findings, the co-owners of the IELTS test have amassed an excessive amount of control over the lives of millions of individuals, which has significant implications for ethical standards. Participants in the IELTS exam were of the opinion that the testing procedure should be made more democratic and considerate of human needs.

Regarding CA, Pepper and Pathak (2008) investigated the perspectives of university test-takers at Southwestern University regarding what constitutes a fair assessment. Their findings showed that the highest priority should be placed on providing frequent feedback, being explicit in assessment administration and grading criteria, and being proactive in assessment practices. This was determined to be the best way to ensure that assessment practices are fair. In addition, Murillo and Hidalgo (2017) investigated the perspectives of Spanish students in primary and secondary education regarding the existence of justice in CA. Their findings demonstrated that fair assessment practices ought to adhere to the principles of equality and equity. Equality requires transparency, objectivity, and an assessment of the material covered in each class. Nonetheless, equity requires qualitative assessment, adaptation to the requirements of test-takers, and the diversification of tests. In addition, Rasooli et al. (2019) made an effort to investigate the perspectives of university students regarding what constitutes a fair assessment of academic performance within the framework of Iranian higher education. According to the findings of the study, the distribution of the results, interpersonal communication among the testing stakeholders, and communication protocols are all essential components of fair assessment practices. Lastly, Fan et al. (2020a) conducted a survey that was focused on hypothetical situations in order to investigate how Chinese university students perceive ethical issues in CA. The results of their investigation showed that the viewpoints of the students and the specialists regarding the circumstance were not compatible with one another. They made the discovery that the students focused their attention most on having different assessment techniques, maintaining secrecy, and communicating about grades.

It is clear from the studies reviewed above that the test-takers' perceptions of ethical requirements in CA in the high school contexts of Iran has remained unexplored. In light of this, the purpose of the current study is to advance our comprehension by drawing on the perspectives of high school test-takers on the ethicality of CA. By fundamentally altering how high school teachers and test-takers conceptualize CA while taking into account test-takers' experiential and perceptual data, it is hoped that the results of this study will help to improve assessment practices. The following research question was proposed to achieve the objectives:

RQ: What are Iranian high school test-takers' perceptions of ethical requirements in language classroom assessment?

Context of the study

Basic education is centralized in Iran and is divided into K-12 education. The Ministry of Education is in charge of the financing and administration of elementary and secondary education. The Ministry of Education, indeed, is responsible for implementing educational policies announced by the government, supervising national exams, educating teachers, developing curricula and learning materials, and building and maintaining schools. Education policies are determined and overseen by the Supreme Council of the Cultural Revolution, Iran's parliament and the cabinet of ministers. After 2012, the education is divided into three levels: primary school (Dabestân), lower secondary school (Dabirestân Dore Aval), and upper secondary school (Dabirestân Dore Dovum). The elementary cycle lasts 6 years. Students attend 24 h of class per week and the curriculum covers Persian studies (e.g., reading, writing, and comprehension), Islamic studies, social studies, science, and mathematics. The lower secondary school cycle lasts 3 years. The curriculum includes history, Arabic, foreign languages, vocational studies, along with the materials taught in the elementary cycle. Based on their grades obtained in the relevant subjects at the end of the lower secondary school cycle, students are eligible to continue their education in the academic or vocational/technical branches of the upper secondary school cycle. The upper secondary school cycle lasts 3 years. It is free at public high schools, but it is not compulsory for students. At this level, students are divided into three fields of the education system: academic (*Nazari*), technical (*Fani Herfei*), and vocational/skills (*Kar-danesh*). Concerning the examination system, it should be noted that examinations are held two times per year, in November and June. Students' scores include two parts: class score and final score. The class score is determined based on students' performance over the course. However, the final score is determined based on students' performance on the tests administered at the end of the course. The grading system ranges from 0 to 20 and the passing score is 10. If the mean of the class score and final score is less than 10, students must repeat the year and may re-take the examination the following year. In grades 10 and 11, teachers are responsible for designing, administering, and grading tests. In grade 12, however, the score class is determined by teachers but the final examinations are designed, administered, and graded by provincial education authorities. Successful students are awarded a Certificate of Diploma.

Table 1 Demographic information of the participants

Participant	Gender	Level	Age
Leila	F	UHS	18
Zoreh	F	UHS	17
Akram	F	UHS	17
Barana	F	UHS	19
Saideh	F	UHS	18
Tamineh	F	UHS	18
Maryam	F	UHS	18
Razieh	F	UHS	19
Masoomeh	F	UHS	18
Fatemeh	F	UHS	20
Narges	F	UHS	17
Mariezeh	F	UHS	18
Samira	F	UHS	18
Neda	F	UHS	19
Rezvan	F	UHS	18

UHS Upper secondary school, *F* Female

Method

Design

In order to carry out the study, the researchers utilized a methodical thematic coding approach (constant-comparative technique). As pointed out by Riazi (2016), this is a method used in grounded theory research that involves classifying and organizing pieces of raw data based on their features and then arranging them in an organized way to create a new theory. In conclusion, this investigation employed a constant- comparative technique to uncover the test takers' views of ethical standards in high school contexts of Iran.

Participants

The present study was run at Zienabieh High School in Borujerd City, Iran. A total of 20 upper high school students were selected using criterion sampling. As Miles and Huberman (1994) note, criterion sampling is a form of purposeful sampling aiming to identify and select information-rich cases in qualitative research. The participants were all female, aged between 17 and 18 years old, and majored in academic fields. The underlying reason for inviting the participants was that they were accessible to the researcher, as well as since they were in grade 12, they had a good understanding of the examination system dominant in the Iranian high schools. The demographic information of the participants is presented in Table 1. To achieve the participants, the first researcher referred to Zienabieh High School. After having a warm greeting with the school officials and the English language teachers, she explained the present study's objectives to them. Given the researcher explanation, they permitted the researcher to run the present study in their school setting. The school principal gave the phone number of the students' parents to the first researcher to contact them and explain the current study's objectives. A written consent form (in Persian) was sent to those parents who agreed to let their children participate willingly in this study. In total, 15 written consent forms were signed and sent back to the researcher. She ensured the school officials, English language teachers, parents, and students that the participation in the present study would be voluntary, the students' responses would remain confidential, and the final findings would be shared with them.

It is noteworthy that the research was monitored and approved by the Ethical Committee Board at Lorestan University in accordance with ethical standards (Code: 35f2/25/112).

Data collection procedures

The researcher used a written reflective statement to collect the required data. The purpose of having a reflective written statement was to give the test-takers an opportunity to think deeply about the ethical requirements that should be followed in CA. As the classroom contexts are primarily similar to the participants across the country, the participants' responses were mingled to establish a broad database for the credibility of the claims. To activate the background knowledge of the participants about ethical requirements in CA, the researchers offered an overview on the issue via a podcast. In particular, the participants were invited to reflect upon the following prompt:

> *Dear Student:*
>
> *You are kindly invited to write a report of your perceptions and experiences with assessment practices wherein you participated as a test-taker. Your report is supposed to be a reflection upon your positive, negative, or neutral perceptions and experiences with the tests and assessment practices administered in the classroom concerning ethical issues. It means that your report is supposed to revolve around ethical issues in classroom assessment. A report with approximately 400–600 words in length would be sufficient.*

The participants were invited to the school office one-by-one. They were asked to reflect upon their diverse perceptions and experiences with different assessment practices in a comfortable environment. The researchers tried not to restrict the participants in expressing their viewpoints by specifying any particular ethical issues. It is worthy to note that the researchers asked the participants to write down their perceptions in their mother tongue (Persian) so that they can express their viewpoints with ease. Later, the researchers recruited two experts in translation to translate the participants' responses into English.

Data analysis procedures

The researcher used a systematic thematic coding procedure to analyze the participants' responses and to extract themes related to ethical issues in CA. The codes with a high degree of co-occurrence (i.e., two or more codes used for the same data) were reframed into broader codes as they represented the same construct (Patton, 2002). For example, 'the use of surprise items' and 'the use of unfamiliar contents' were predominantly coded together and they were, thus, subsumed under the broader code of 'avoidance of unfamiliar contents and surprise items'. The extracted themes represented a collection of associated codes according to a code list and frequency counts per code. As Guest and

McLellan (2003) note, co-occurrence is an appropriate way to clarify the relations among some conceptual categories. The usefulness of co-occurrence lies in the fact that "categorizing is never just an end in itself. Its goals are often the discovery and ordering of ideas and themes; and the storing of growing understandings, the linking of ideas to data, cross-referencing, sorting and clarifying" (Richards and Richards, 1995, p. 80 as cited in Cheng & DeLuca, 2011). In other words, the co-occurrence allowed the researcher to give more than one theme to the participants' statements. For example, if a participant commented that teachers should consider multiple assessment methods with familiar content, these data gained double codes (i.e., use of alternative assessment methods and avoidance of unfamiliar contents and surprise items). The advantage of data analysis through co-occurrence was letting the researcher find the relationships between themes and test-taking experiences. The increased frequency of instances of co-occurrence suggested a more significant association between the participants' perceptions and experiences. Co-occurrence percentages were calculated based on the ratio of co-occurrence frequency of the extracted themes. The results of co-occurrence analysis are presented in Appendix.

The researcher ensured the reliability and credibility of the findings. Concerning reliability, the researcher recruited two coding analysts to code the collected data, and the inter-rater reliability of their coding was 0.92. They tried to reach a consensus by discussing coding when a disagreement emerged. Regarding credibility, the researcher used a member checking strategy. In doing so, the researcher gave a copy of the extracted codes and themes to five participants to confirm or add more comments. In general, they confirmed the results and interpretations.

Results and discussion

Having meticulously analyzed the data, the researchers reached. The findings yielded two overarching categories, including *do no harm* (e.g., establishing a supervision group, considering test-takers' individual differences, keeping test results confidential, and turning back test sheets with feedback) and *avoid score pollution* (e.g., using additional knowledge sources, using alternative assessment methods, clarifying grading criteria, avoiding unfamiliar contents and surprise items) (see Fig. 1). They are detailed below.

Do no harm

Establishing a supervision group

"Establishing a supervision group" was the first theme related to do no harm that emerged from the database. The participants maintained that to ascertain the ethicality requirements of assessment practices, a supervision group, including some experts in L2 assessment should be established. The following excerpt shows this:

> *"A supervision group is missing in high schools. This group can supervise if assessment practices are ethical by securitizing test design, test administration, test grading, and test results interpretation. If there is a problem or shortcoming, they can provide the required pieces of advice" (Leila, May 18, 2021).*

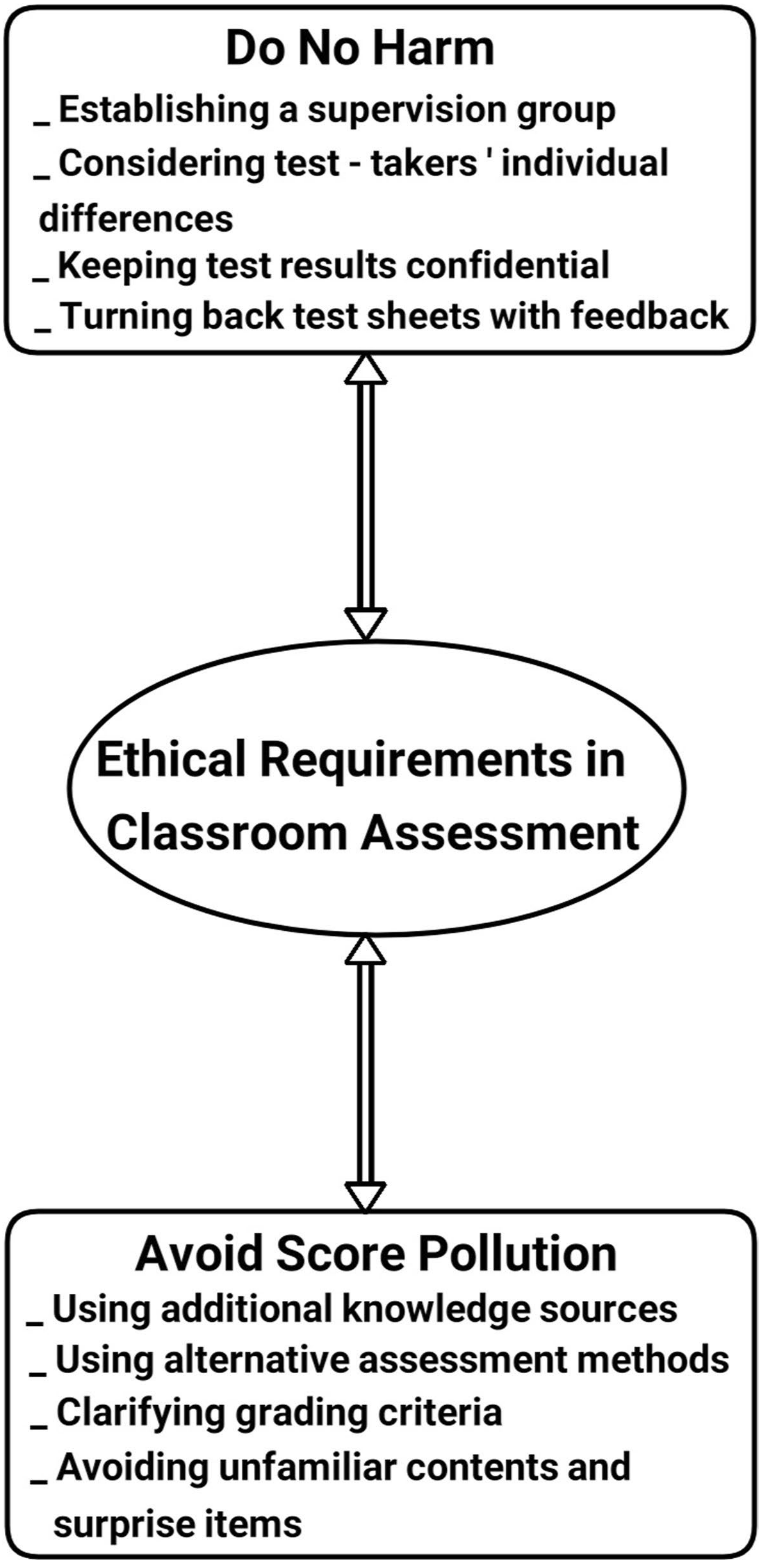
Do No Harm
_ Establishing a supervision group
_ Considering test - takers ' individual differences
_ Keeping test results confidential
_ Turning back test sheets with feedback
Ethical Requirements in Classroom Assessment
Avoid Score Pollution
_ Using additional knowledge sources
_ Using alternative assessment methods
_ Clarifying grading criteria
_ Avoiding unfamiliar contents and surprise items

Fig. 1 Ethical requirements in classroom assessment

Moreover, the participants stressed that establishing a supervision group could be helpful to examine whether tests are appropriate for the intended purposes. For this, one of the participants opined:

> *"Given the fact that most test-takers, parents, and other testing stakeholders are not familiar with the basic feature of a quality test, they may not have any idea if ethical requirements are met in testing practices. Therefore, this supervision group can report if testing stakeholders right has been preserved in assessment practices" (Barana, May 18, 2021).*

The above statements are very indicative of 'establishment of a supervision group'. The findings can be explained from this view that when a supervision group is established, different testing stakeholders may feel assured that assessment processes have been followed correctly (Lynch, 2001). In line with Shohamy (2001), it may be argued that this supervision group can support test-takers' rights by mirroring their voices and concerns to teachers and school officials. As there exists an unbalanced distribution of power between teachers and test-takers "either politically or educationally to control the educational system and to inject specific priorities to the society" (Shohamy, 1997, p. 345), establishing this supervision group may lead to a more democratic environment wherein test-takers' voice and concerns are heard. Therefore, according to Shohamy (2001b), establishing this supervision group set the stage for a balanced distribution of power by creating a channel through which teachers and test-takers can exchange their views. In this way, they can sort out the potential assessment problems.

Considering test-takers' individual differences

The next theme extracted from the participants' responses was "considering test-takers' individual differences". This was related to do no harm category. The participants maintained that to make assessment practices ethical, the test-takers' individual differences, such as learning styles, age, gender, first language, and ethnicity should be taken into account. In this regard, one of the participants commented:

> *"It is a reality that test-takers differ from each other. So, teachers need to be aware of this in the design, administration, and development of assessment practices. For example, I am left-handed, and unfortunately, most of the time, there is no appropriate chair for me during test administration" (Maryam, May 19, 2021).*

In corroborating with the previous statements, one of the participants quoted:

> *"I think that it is not ethical if teachers leave test-takers' individual differences unnoticed. For example, test contents should not unduly privilege a group of test-takers. In one of my experiences, some of my classmates got high scores on a reading comprehension test because the test contents were in harmony with their cultural background" (Zohreh, May 22, 2021).*

The above statements indicated that assessment practices are perceived as ethical if test-takers' individual differences are given attention by teachers. One possible explanation of the study's findings is that since not every student learns in the same ways, they prefer to demonstrate their learning differently (Appel & Wood, 2016). In this process, teachers should use different assessment methods so that student can show their abilities. For example, as an essential individual characteristic, the study's findings of Breland et al. (2004) evidenced that gender was a significant predictor of essay writing performance where female learners tended to obtain higher scores than male learners. Another possible explanation of the study's findings, as Crossley and Kim (2019) note, is that when test-takers' individual differences are not considered in assessment practices, the test-takers' scores may be polluted by irrelevant constructs. For instance, some test-takers show their learning best by seeing, hearing, touching, and reading. The study's findings receive support from Lynch (2001), asserting that issues of age, gender, ethnicity, ideological beliefs, and first culture should be considered in assessment practices.

Keeping test results confidential

"Keeping test results confidential" was another theme related to do no harm catching the participants' attention. In this respect, one of the participants commented:

> *"It is not ethical to publicize test-takers' scores in front of the classroom. Instead of announcing test-takers' scores, teachers can give them directly to test-takers. In this way, the test-takers who get low scores, are not got ridiculed by other test-takers and they do not become ashamed of their scores." (Fatemeh, May 18, 2021).*

Corroborating with the previous statements, the participants pinpointed that the confidentiality of scores is one of the test-takers' undeniable rights. The following statement clearly shows this:

> *"If teachers want to respect my privacy right, they should keep my scores confidential. Other test-takers are not legitimate of being informed about my scores" (Narges, May 14, 2021).*

The above statements lend credence to this view that keeping test results is one of the crucial aspects of ethics in CA. The findings can be explained from this view that teachers need to admit that test-takers, as human beings, have the privacy rights in assessment practices (Fan et al. 2020a; Shohamy, 2013). By publicizing test results, they violate test-takers' privacy rights. Another possible explanation of the study's findings is that test-takers can manage their studies and learn best from their performance on tests, if their scores are kept confidential in the classroom context (Davis, 2010; Moore, 2005). The study's findings support Worthen et al. (1998), concluding that test results should not be revealed to anyone who does not have a legitimate need to know the scores. The study's findings also are consistent with the previous studies (Brookhart & Nitko, 2008; Fant et al. 2020a; Murrillo and Hidalgo, 2017), reporting that one of the ethical requirements of assessment practices is keeping test results confidential.

Turning back test sheets with feedback

The last theme related to do no harm extracted for the database was 'turning back test sheets with feedback'. The participants held that it is ethical to turn back test-takers' test sheets with feedback. In support of this, one of the respondents expressed:

> *"I think that it is not ethical that my teacher does not return my test sheets back. Due to this problem, I cannot find out what my strengths and weaknesses have been*

> *on tests" (Razieh, May 18, 2021).*

Additionally, one of the participants quoted:

> *"By providing feedback on my test performance, I can learn from my mistakes and re-structure the linguistic structures in my mind. But, unfortunately, my teacher never turns back test sheets" (Maryam, May 18, 2021).*

The above statements clearly showed that one of the ethical requirements of assessment practices is turning back test sheets with feedback. The study's findings can be explained from this view that by turning back student's test sheets with feedback, positive wash-back effects are created. In this way, test-takers can get a better picture of their performance and can strengthen their strong points and remedy their weak points (Carless, 2006; Fan et al. 2020a). Another possible explanation of the findings is that when the feedback on test-takers' performance is transparent (i.e., the feedback is easy to decipher its meaning and intention), consistent (i.e., feedback is in congruent with the previous comments), and justification (i.e., feedback is logical), it can be motivating for test-takers to continue learning (Rasooli et al, 2018). The study's findings accord with the previous study (Carelss, 2006; Fan et al. 2020a; Lizzio & Wilson, 2008; Rasooli et al., 2018), indicating that providing feedback in test sheets makes assessment practices be perceived as fair by test-takers.

Avoid score pollution

Using additional knowledge sources

One of the dominant themes related to avoid score pollution gaining noticeable co-occurrence in the database was "using additional knowledge sources". This theme indicates that teachers need to benefit from additional knowledge sources such as colleagues, test-takers, and parents in the design, administration, and grading of assessment practices. In this regard, one of the participants commented:

> *"To improve the quality of assessment practices, I think, teachers need to seek other knowledge sources. For example, when teachers consult with their colleagues, they can take advantage of their experiences and vantage points to implement assessment practices. In this way, they can ensure that quality requirements, including ethicality are met well" (Rezavan, May 10, 2022).*

The participants stressed that when teachers use other knowledge sources, they can implement assessment practices in line with test-takers' needs and characteristics. In this respect, one of the participants quoted:

> *"My experiences have demonstrated that when teachers use test-takers' views to implement assessment practices, they are more positive and constructive for student learning. For example, our English teachers usually ask test-takers' views and concerns in implementing assessment practices. This collective knowledge makes assessment practices be tailored to test-takers' needs and wants. Therefore, they would be more ethical" (Samira, May 22, 2022).*

The above statements evidenced that teachers need to consult with other knowledge sources to meet ethical requirements in assessment practices. The study's findings can be explained with the words of Shohamy (2001b), arguing that since "the knowledge of any tester is incomplete", testers' knowledge should be "negotiated, challenged, and appropriated" (p. 132). Further, along with Dimova and Kling (2018), another possible explanation of the study's findings is that if a dynamic, cooperative climate is created among all testing stakeholders, it is more likely to achieve higher reliability and validity, and in consequence, to make assessment practices more ethical. Additionally, the study's findings may be explained from this view that when test-takers' voices are heard by teachers, test-takers may engage more actively in assessment practices, they may get a better understanding of the complexity of ethical issues, and they may have chances to discuss teacher's feedback (Flores et al., 2015; Murrillo and Hidalgo, 2017; Rasooli et al., 2018). The study's findings lend support to the previous studies (Hamp-Lyons, 2000; Safari, 2016; Tahmasebi & Yamini, 2013) documented that although test-takers' voices are the least heard in assessment practices, they should be considered as a significant knowledge source to administer ethical assessment practices.

Use of alternative assessment methods

Another theme related to do no harm that received noticeable attention from the participants was "using alternative assessment methods". In contrast to conventional assessments, alternative assessments usually necessitate that test-takers to think about their general learning to decide what knowledge and abilities they must employ to address a given assessment task (Janisch et al., 2007). The participants pinpointed that teachers should not restrict assessing test-takers' learning to just one kind of assessment method. Teachers should use alternative assessment methods, such as peer-assessment and portfolio assessment to reach a comprehensive understanding of student learning. In support of this, one of the participants stated:

> *"In the elementary school cycle, we had good experiences with alternative assessment methods like portfolio assessment. Unfortunately, it is rarely practiced in our classroom. Since our English teacher usually uses one assessment method, we cannot show our abilities well. I think this is not fair" (Akram, May 14, 2022).*

Another point that supported using different alternative assessment methods was related to the fact that test-takers with different cognitive learning styles may perform differently on different assessment practices. The following excerpt clearly shows this:

> *"I feel that administering alternative assessment methods can be highly promising and motivating for us. The reason is that we test-takers learn in different ways and show our learning in different ways too. For example, I am good at answering multiple-choice tests but I am terrible with essay tests. Hence, it is more ethical if our teacher uses different assessment methods" (Razieh, May 19, 2022).*

The above statements documented that 'use of alternative assessment methods' is an integral component to administer ethical assessment practices. The study's findings may be explained from this view that alternative assessment methods should be practiced since every assessment method enjoys some advantages, as well as suffer from some disadvantages. For example, Green and Johnson (2010) demonstrate that though multiple-choice tests are helpful to cover a wide variety of content areas, they do not lead themselves to measure student creativity as a higher-level cognitive skill. Further, another possible explanation of the study's findings is that using alternative assessment methods, such as portfolio assessment makes assessment practices more process-oriented (Fan et al. 2020a). In this way, teachers may get a better picture of student learning. The study's findings lend support to Tierney (2016), arguing that to make assessment practices ethical, teachers should provide test-takers with "multiple, varied, equitable, and meaningful opportunities to demonstrate their learning" (p. 6). The study's findings also accord with the previous studies (e.g., Brookhart & Nitko, 2008; Gronlund, 2003; Ishihara & Chiba, 2014; Rasooli et al., 2018), recommending that no single test should be used because no single test can adequately measure student learning.

Clarifying grading criteria

"Clarity of grading criteria" was the next theme extracted from the database. This was pertinent to do no harm category. Although the participants highlighted the importance of the clarity of grading criteria, they blamed that it is usually missing in assessment practices. In this regard, one of the respondents quoted:

> *"I feel that it is ethical to keep test-takers informed about grading criteria. By doing so, we, test-takers, get informed about the weight of learning materials, and consequently, we can manage our studies better. Despite this significance, my teacher does not clarify grading criteria" (Masoomeh, May 18, 2021).*

Resonated with the previous statement, the participants stressed that by clarifying grading criteria, they could get a better understanding of learning objectives. In support of this, one of the participants remarked:

> *"Since the grading criteria are not clear to me, I do not know what the learning objectives and my teachers' expectations are. Therefore, I do not usually get my desired scores and, consequently, I lose my motivation to continue learning" (Neda, May 18, 2021).*

The above statements evidenced that there should be clarity in the grading criteria in assessment practices in the classroom. The participants' perceptions clearly show that teachers are responsible for clarifying grading criteria to test-takers. One possible explanation of the study's findings, as Tierney (2013) argues, is that test-takers' opportunities to learn and opportunities to demonstrate learning increase by providing them with clear learning and assessment expectations. Clarifying the grading criteria can be of great help for test-takers to perceive the learning objectives and teachers' performance expectations (Alm & Colnerud, 2015; Camilli, 2006; Gipps & Stobart, 2009). Another possible explanation of the study's findings is that as test-takers know the grading criteria, they manage their studies better to get higher scores. This, in turn, may increase their motivation for continued learning and improvement (Brookhart & Nitko, 2008). The study's findings are congruent with the previous studies (Fan et al. 2020a; Pepper & Pathak, 2008; Rasooli et al., 2019; Tierney, 2016), demonstrating that with presence of a clarity in grading criteria, test-takers perceived assessment practices as fair.

Avoiding unfamiliar contents and surprise items

The other theme that emerged from the participants' responses was "avoiding unfamiliar contents and surprise items". The participants blamed that sometimes tests contents do not mirror the educational materials taught during the course, which jeopardizes the ethicality of assessment practices.

> *"I think testing practices are not ethical when tests contents are not familiar to all test-takers. When tests include items that have not been taught during the course, I cannot show my real abilities. For example, in one of my experiences, I took a grammar test containing some unfamiliar grammatical structures. However, I did not have the right not to take the test and complain about its unfamiliar contents. Unfortunately, I failed the test" (Fahimeh, May 18, 2021).*

Additionally, the participants blamed that sometimes there is a lack of transparency in the curriculum and the educational materials presented by teachers. This leads to surprise items in tests. In support of this, one of the test-takers stated:

> *"The lack of transparency in educational materials is a big problem. It is not fair that sometimes high school teachers do not verify well the intended curriculum and educational materials. This lack of transparency makes test-takers encounter unfamiliar test items or so-called 'surprise items'" (Saideh, May 18, 2021).*

The above statements evidenced that one of the issues that may jeopardize the ethicality of assessment practices is using unfamiliar contents and surprise items. As the findings documented, the participants blamed that sometimes there is a disparity between test items and the educational materials worked during the course. One possible explanation of the findings may be attributed to the testing power and testing culture practiced in the Iranian high school contexts. In the Iranian testing culture, due to exercising bureaucratic imposition mandated by high schools, test-takers may not have the power to criticize the test contents and test methods, let alone refuse to take them (Safari, 2016). This argument can be supported by Shohamy (2013), claiming that sometimes test-takers have to comply with any demands that test-makers and test-users make, since passing the tests are considered necessary requisites to receive a societal membership, to be admitted in tertiary education, or to get a job. That is, as passing a test is the license for achieving benefits, people do anything to succeed on tests, "even if the demands of the tests are perceived as detached from their perceptions of 'true' and 'real' knowledge" (Shohamy, 2013, p. 5). However, in light of the findings, it can be argued that if teachers aim to make assessment practices ethical, they should not include any surprise items whose content has not been covered during the course. The study's findings are compatible with the previous studies (Farhady & Hedayati, 2009; Safari, 2016; Tahmasebi & Yamini, 2013), indicating that although sometimes test-takers have to comply with tests contents, purposes, and demands due to life-long decisions made with reference to test results, test-makers should avoid including unfamiliar contents and surprise items. Otherwise, assessment practices may not be perceived as ethical by test-takers.

Conclusion and implications

This study purported to explore test-takers' perceptions about the status of ethics in CA in the context of Iranian high schools. As reported and discussed in the preceding section, The findings yielded two overarching categories, including *do no harm* (e.g., establishing a supervision group, considering test-takers' individual differences, keeping test results confidential, and turning back test sheets with feedback) and *avoid score pollution* (e.g., using additional knowledge sources, using alternative assessment methods, clarifying grading criteria, avoiding unfamiliar contents and surprise items). The findings documented that concerning assessment practices, the test-takers are not satisfied with the current status quo. This dissatisfaction may be attributed to the ignorance of ethical requirements by teachers. If teachers aim to alleviate the current status quo, there is an urgent call for fundamental reforms in assessment practices in terms of ethical considerations. For example, one of the essential changes is starting a gradual transition from the current assessment practices to ethical assessment programs in the Iranian teacher education courses. These teacher education courses can provide an invaluable opportunity for teachers to become familiar with the tenets of ethics.

In light of the study's findings, some implications are offered for different testing stakeholders. The first implication is that public and academic awareness should be raised about the study's findings. It should be publicized that assessment practices should be very sensitive to ethical aspects (Shohamy, 2013). By doing so, it is likely to improve the testing contexts and assessment practices in high school contexts. The next implication is that teachers should practice alternative assessment methods. By practicing alternative assessment methods, teachers can reach a more valid response to the ethicality need as opposed to the summative-only type of assessment in which one dimension of test-takers' language ability is tested at one specific time (Rasooli et al., 2018). Another implication is that additional knowledge sources such as test-takers' and parents' voices and views should be incorporated in assessment practices. The inclusion of test-takers' and parents' concerns in assessment practices make them admit test results and decisions made based on test results. They would not probably complain that their abilities have been evaluated unfairly. However, as Rasooli et al. (2018) propose, to meet this purpose, "teachers can build a constructive classroom environment through respectful relationship, listening to test-takers, and enacting the do no harm principle" (p. 171). Another implication is related to the use and interpretation of test results in light of test-takers' individual differences. As there are limitations and uncertainties with test scores, they should be used and interpreted carefully in light of test-takers' individual differences (e.g., age, gender, ethnicity, L1 and culture, and L2 proficiency). The other implication is that teachers need to offer constructive feedback on test-takers' performance and turn back test sheets. By doing so, test-takers get a clear picture of their learning. The following implication is that grading criteria should be transparent, consistent, and logical. As Rasooli et al. (2018) stress, to promote ethicality in assessment practices, a cycle of explanation of grading criteria, their justifications, and their consistent applications should be implemented in the classroom. The final implication is that, as assessment ethicality literacy is an essential component of assessment literacy, pre-service and in-service teacher training courses should be devised so that teachers reach a good body of knowledge of the multidimensional view of ethicality in assessment practices. In this way, it may be ascertained that teachers can understand the significance of ethicality in assessment practices in the classroom.

Here, considering the limitations imposed on the present study, some suggestions for further research are presented. First, as the sample of the present study was limited to only twenty female high school test-takers, larger samples with male gender can increase the generalizability of the present study's findings. Second, since this study was limited to the high school contexts, further research is needed to explore the ethical requirements in high-stakes and low-stakes tests in Iran from test-takers', parents', and education officials' perceptions. Third, because this study focused on the ethical requirements, future studies can empirically scrutinize the kind and amount of wash-back to the actual classes, when ethical requirements are met in assessment practices. Fourth, more studies are needed to investigate EFL teachers and test-takers' perceptions of the effects of (un)fair feedback on achieving ethicality in CA. Fifth, more research needs to explore test-takers' and teachers' perceptions about ethicality in diverse sociopolitical and cultural contexts such as democratic and undemocratic and individualistic and collectivist societies. Sixth, future research can probe into teachers' and test-takers' perceptions about ethicality in teaching practices and its effects on assessment practices. Finally, a further study is needed to disclose the possible effects of test-takers' individual differences on uses and interpretations of test results in the Iranian testing programs from testing stakeholders' perceptions.

The last point is that since, to the best of our knowledge, this is the single most comprehensive study that has been done in an Iranian high school context with a particular group of high school test-takers over a specific moment concerning ethical considerations in CA, enough care should be exercised prior to generalizing the findings to the whole system of the Iranian education system.

Appendix

Table 2 Theme frequencies

Themes	Frequency
Use of communal knowledge	42
Use of alternative assessment methods	35
Attention to students' individual differences	30
Establishment of a supervision group	35
Avoidance of unfamiliar contents and surprise items	28
Keeping test results confidential	25
Clarity of grading criteria	22
Turning back test sheets with feedback	21

Abbreviation

CA	Classroom assessment

Acknowledgements
Not applicable.

Authors' contributions
All authors have made substantial contributions to conception and design, acquisition of data, analysis and interpretation of data, and writing the manuscript. The author(s) read and approved the final manuscript.

Authors' information

Zeinab Azizi holds a Ph.D. in TEFL. In 2020, Azizi joined the Department of English, Ayatollah Borujerdi University, as a faculty member. Her areas of interest include online education, teacher education, and sociocultural studies. She has published some papers in domestic and international journals.
Seyed M. Ismail is an assistant professor at Prince Sattam Bin Abdulaziz University, Saudi Arabia. His research interests are teaching and learning, testing and educational strategies. He published many papers in different journals.

Funding

This study is supported via funding from Prince Sattam Bin Abdulaziz University Project Number (PSAU 2023 /R/1444).

Availability of data and materials

The authors state that the data supporting the findings of this study are available within the article.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 23 February 2023 Accepted: 13 April 2023
Published online: 08 June 2023

References

Ahmadi, R. (2021). Students' perceptions of student voice in assessment within the context of Iran: The dynamics of culture, power relations, and student knowledge. *Higher Education Research & Development*. https://doi.org/10.1080/07294360.2021.1882401

Airasian, P. (2005). *Classroom assessment - concepts and applications* (5th ed.). McGraw-Hill.

Alm, F., & Colnerud, G. (2015). Teachers' experiences of unfair grading. *Educational Assessment,20*, 132–150. https://doi.org/10.1080/10627197.2015.1028620

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly,13*, 55–71. https://doi.org/10.1080/15434303.2015.1126718

Association of Language Testers in Europe. (2001). *Principles of good practice for ALTE examinations*. Retrieved March 30, 2009 from http://www.alte.org/cop/principles.php

Azizi, Z. (2022). Fairness in assessment practices in online education: Iranian University English teachers' perceptions. *Language Testing in Asia,12*(1), 1–17. https://doi.org/10.1186/s40468-022-00164-7

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Breland, H., Lee, Y., Najarian, M., & Muraki, E. (2004). An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups. *TOEFL researcher reports report 76*. ETS.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Pearson College Division.

Brown, G. T., & Harris, L. R. (Eds.). (2016). *Handbook of human and social conditions in assessment*. Routledge.

Brown, M., McNamara, G., O'Brien, S., Skerritt, C., O'Hara, J., Faddar, J., ... & Kurum, G. (2020). Parent and student voice in evaluation and planning in schools. *Improving Schools*, *23*(1), 85–102. https://doi.org/10.1177/1365480219895167

Camilli, G. (2006). Test fairness. Educational Measurement, 4, 221-256. https://doi.org/10.1186/s40468-023-00235-3

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education,31*, 219–233. https://doi.org/10.1080/03075070600572132

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing,25*(1), 12–25. https://doi.org/10.1177/0265532207083743

Cheng, L., & DeLuca, Ch. (2011). Voices from test-Takers: Further evidence for language assessment validation and use. *Educational Assessment,16*(2), 104–122. https://doi.org/10.1080/10627197.2011.584042

Cohen, J. (2006). Social, emotional, ethical and academic education: Creating a climate for learning, participation in democracy and well-being. *Harvard Educational Review, 76*(2), 201–237. https://doi.org/10.17763/haer.76.2.j44854x1524644vn

Crossley, S. A., & Kim, Y. (2019). Text integration and speaking proficiency: Linguistic, individual differences, and strategy use Considerations. *Language Assessment Quarterly,12*(3), 1–20. https://doi.org/10.1080/15434303.2019.1628239

Davies, A. (1997). Demands of being professional in language testing. *Language Testing,14*, 328–339. https://doi.org/10.1177/026553229701400309

Davies, A. (2008). *Assessing academic English testing English proficiency 1950–89: The IELTS solution.* Cambridge University Press.

Dimova, S., & Kling, J. (2018). Assessing English-medium instruction lecturer language proficiency across disciplines. *TESOL Quarterly,52*(3), 634–656. https://doi.org/10.1002/tesq.454

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.

Fan, X., Johnson, R., & Liu., X. (2017). Chinese university professors' perceptions about ethical issues in classroom assessment practices. *New Waves Educational Research & Development,20*(2), 1–19.

Fan, X., Johnson, R., Liu, X., & Zhang, T. (2019). A comparative study of pre-service teachers' views on ethical issues in classroom assessment in China and the United States. *Frontiers of Education in China,14*(2), 309–332. https://doi.org/10.1007/s11516-019-0015-7

Fan, X., Johnson, R., Liu, X., & Gao, R. (2020). College students' views of ethical issues in classroom assessment in Chinese higher education. *Studies in Higher Education,45*(8), 1–15. https://doi.org/10.1080/03075079.2020.1732908

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics,29*, 132–141. https://doi.org/10.1017/S0267190509090114

Flores, M. A., Veiga Simão, A. M., Barros, A., & Pereira, D. (2015). Perceptions of effectiveness, fairness and feedback of assessment methods: A study in higher education. *Studies in Higher Education,40*, 1523–1534. https://doi.org/10.1080/03075079.2014.881348

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith, & J. Joy Cumming (Eds.). *Educational assessment in the 21st century* (pp. 105–118). Springer.

Gitsaki, C., & Robby, M. A. (2018). Benefits of language assessment. *The TESOL Encyclopedia of English Language Teaching*, 1–6.

Green, S., Johnson, R., Kim, D., & Pope, N. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education,23*(7), 999–1011. https://doi.org/10.1016/j.tate.2006.04.042

Green, S. K., & Johnson. R. L. (2010). *Assessment is essential*. McGraw-Hill Higher Education.

Gronlund, N. (2003). *Assessment of student achievement* (7th ed.). Allyn and Bacon.

Guest, G., & McLellan, E. (2003). Distinguishing the trees from the forest: Applying cluster analysis to thematic qualitative data. *Field Methods,15*, 186–201. https://doi.org/10.1177/1525822X03015002005

Hall, G. (2009). International English language testing: A critical response. *ELT Journal,64*(3), 321–328. https://doi.org/10.1093/elt/ccp054

Hamid, M. O. (2014). World Englishes in international proficiency tests. *World Englishes,33*(2), 263–277.

Hamid, M. O., & Hoang, N. T. H. (2018). Humanizing language testing. *TESL-EJ,22*(1), 25–37.

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language Testing in Asia,2019*, 9–16. https://doi.org/10.1186/s40468-019-0092-9

Hamid, M. O. (2016). Policies of global English tests: Test-takers' perspectives on the IELTS retake policy. *Discourse: Studies in the Cultural Politics of Education, 37*(3), 472–487.https://doi.org/10.1080/01596306.2015.1061978

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System,28*(4), 579–591. https://doi.org/10.1016/S0346-251X(00)00039-7

Homan, R. (1991). The ethics of social research. Addison-Wesley Longman Limited.

Hyatt, D. (2013). Stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *Journal of Further and Higher Education,37*(6), 844–863. https://doi.org/10.1080/0309877X.2012.684043

International Language Testing Association. (2000, March). *Code of ethics for ITLA*. Retrieved October 28, 2007, from http://www.iltaonline.com/code.pdf

International Language Testing Association. (2005, July). ILTA: Draft code of practice: Version 3. Retrieved October 28, 2007, from http://www.iltaonline.com/CoP_3.1.htm

Ishihara, N., & Chiba, A. (2014). Teacher-based or interactional? Exploring assessments for children's pragmatic development. Iranian Journal of language testing, 4(1), 84–112.

Janisch, C., Liu, X., & Akrofi, A. (2007). Implementing alternative assessment: opportunities and obstacles. In *the educational forum* (Vol. 71, No. 3, pp. 221–230). Taylor & Francis Group.

JCSEE. (2003). *The student evaluation standards*. Corwin.

JCSEE. (2015). *The student evaluation standards* (2nd ed.). Corwin.

Johnson, R. L., Green, S. K., Kim, D. H., & Pope, N. S. (2008). Educational leaders' perceptions about ethical practices in student evaluation. *American Journal of Evaluation,29*(4), 520–530. https://doi.org/10.1177/1098214008322803

Liu, J., Johnson, R., & Fan, X. (2016). A comparative study of Chinese and United States pre-service teachers' perceptions about ethical issues in classroom assessment. *Studies in Educational Evaluation,48*, 57–66. https://doi.org/10.1016/j.stueduc.2016.01.002

Lizzio, A., & Wilson, K. (2008). Feedback on assessment: Students' perceptions of quality and effectiveness. *Assessment & Evaluation in Higher Education*, 33, 263–275. https://doi.org/10.1080/02602930701292548.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language testing,*

Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *Language Assessment Quarterly,1*(2–3), 123–135. https://doi.org/10.1080/15434303.2004.9671780

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

Moore, H. M**.** (2005). *Identifying the target population: a genealogy of policy-making for English as a second language (ESL) in Australian schools*. Unpublished Ph.D. dissertation, University of Toronto, Toronto: Canada.

Murchan, D., & Siddiq, F. (2021). A call to action: A systematic review of ethical and regulatory issues in using process data in educational assessment. *Large-Scale Assessment in Education,9*, 1–27. https://doi.org/10.1186/s40536-021-00115-3

Murillo, F. J., & Hidalgo, N. (2017). Students' conceptions about a fair assessment of their learning. Studies in Educational Evaluation, 53, 10-16. https://doi.org/10.1016/j.stueduc.2017.01.001

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly,8*(2), 146–160. https://doi.org/10.1080/15434303.2011.564698

Oosterhof, A. (2009). *Developing and using classroom assessments* (4th ed.). Pearson.

Ory, J. C., & Ryan, K. E. (1993). *Tips for improving testing and grading* (Vol. 4). Sage.

Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Sage Publications, Inc.

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal,35*(2), 1–10. https://doi.org/10.1093/elt/ccz006

Pepper, M., & Pathak, S. (2008). Classroom contribution: what do students perceive as fair assessment? *Journal of Education for Business*, 360–368. https://doi.org/10.3200/JOEB.83.6.360-368.

Pope, N. S. (2006). *Do no harm to whom?* (pp. 25–31). South Atlantic Philosophy of Education Society Yearbook. 25–31. South Atlantic Philosophy of Education Society. https://drive.google.com/file/d/0B2lRXFWcJbkldnRMVURackJNUGc/.

Pope, N., Green, S. K., Johnson, R. L., & Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment. *Teaching and Teacher Education,25*(5), 778–782. https://doi.org/10.1016/j.tate.2008.11.013

Pope, N. S. (2006a). Do no harm to whom? An examination of ethics and assessment, *South Atlantic Philosophy of Education Society Yearbook*, 25-31

Popham, W. J. (1991). Appropriateness of teachers' test preparation practices. *Educational Measurement: Issues and Practice,10*(4), 12–15. https://doi.org/10.1111/j.1745-3992.1991.tb00211.x

Popham, W. J. (2000). Big change questions. Should large-scale assessment be used for accountability? Answer: Depends on the assessment, Silly! *Journal of Educational Change*, *1*(3), 283–289. DOI:https://doi.org/10.1023/A:1010054525759

Popham, W. J. (2017). *Classroom assessment: what teachers need to know* (8th ed.). Pearson.

Rasooli, A., Zandi, H., & DeLucab, Ch. (2018). Re-conceptualizing classroom assessment fairness: A systematic metaethnography of assessment literature and beyond. *Studies in Educational Evaluation,56*, 164–181. https://doi.org/10.1016/j.stueduc.2017.12.008

Rasooli, A., DeLuca, Ch., Rasegh, A., & Fathi, S. (2019). Students' critical incidents of fairness in classroom assessment: An empirical study. *Social Psychology of Education,22*, 701–722. https://doi.org/10.1007/s11218-019-09491-9

Rezai, A. (2022). Fairness in classroom assessment: Development and validation of a questionnaire. *Language Testing in Asia,12*(1), 1–27. https://doi.org/10.1186/s40468-022-00162-9

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*(1), 1–20. https://doi.org/10.1186/s40468-022-00157-6

Riazi, A. M. (2016). The Routledge encyclopedia of research methods in applied linguistics. Routledge.

Richards, T., & Richards, L. (1995). Using hierarchical categories in qualitative data analysis. Computer-aided Qualitative Data Analysis: Theory, Methods, and Practice, 80-95.

Safari, P. (2016). Reconsideration of language assessment is a MUST for democratic testing in the educational system of Iran. *Interchange,47*(3), 267–296. https://doi.org/10.1007/s10780-016-9276-8

Sax, G. (1974). *Principles of educational measurement and evaluation*. Wadsworth.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing,14*(3), 340–349. https://doi.org/10.1177/026553229701400310

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

Shohamy, E. (2001b). Democratic assessment as an alternative. Language Testing, 18(4), 373-391. https://doi.org/10.1177/026553220101800404

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

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication,13*(2), 1–12.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice,8*(2), 5–14. https://doi.org/10.1111/j.1745-3992.1989.tb00315.x

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), 103–126. https://www.sid.ir/en/journal/ViewPaper.aspx?id=310596

Taylor, C. S., & Nolen, S. B. (2005). *Classroom assessment: Supporting teaching and learning in real classrooms*. Prentice Hall.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. Language Testing 30(3), 403-412. https://doi.org/10.1177/0265532213480338

Tierney, R. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 125–144). Sage.

Tierney, R. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation,43*, 55–69. https://doi.org/10.1016/j.stueduc.2013.12.003

Tierney, R. (2016). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1–6). Springer.

Uysal, H. H. (2009). A critical review of the IELTS writing test. *ELT Journal,64*(3), 314–320. https://doi.org/10.1093/elt/ccp026

Waugh, K., & Gronlund, N. E. (2013). *Assessment of students achievement* (10th ed.).

Weir, C. J. (2005). *Language testing and validation*. Palgrave Macmillan.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks. R. (1998). *Measurement and assessment in the schools* (2nd ed.). Allyn & Bacon of Pearson.

Xiao, Y., & Carless, D. R. (2013). Illustrating students' perceptions of English language assessment: Voices from China. *RELC Journal,44*(3), 90–105. https://doi.org/10.1177/0033688213500595

Publisher's Note

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