Content list available at http://ijltr.urmia.ac.ir

Iranian Journal of Language Teaching Research

ORIGINAL ARTICLE

Urmia University

Reconceptualization of Test Fairness Model: A Grounded Theory Approach

Shima Beheshti [a], Mohammad Ahmadi Safa [a, *]

[a] *Bu-Ali Sina University, Iran*

ABSTRACT

The indefinite nature of test fairness and different interpretations and definitions of the concept have stirred a lot of controversy over the years, necessitating the reconceptualization of the concept. On this basis, this study aimed to explore the empirical validity of Kunnan's (2008) Test Fairness Framework (TFF) and revisit the established test fairness conceptualization following the principles of grounded theory. To this end, 10 university lecturers of TEFL, 20 high school English language teachers, 15 PhD students in TEFL, and 15 MA students in TEFL participated in open-ended and semi- structured interviews. Following grounded theory rubrics, the researchers read, codified, and analyzed the obtained interview data. Simultaneously, memos were written, comparisons were drawn, possibilities were seen, and robust categories were developed through theoretical sampling. This process continued iteratively until the categories saturated. Next, the categories were juxtaposed and compared to see how they fit together and finally several major categories emerged accordingly. The opinions were diagrammed and a visual image of the categories and their relevant scope, power, and associations were represented to construct a theoretical logic. The new hierarchy of test fairness categories became discernible as the interviewees named distinct characteristics for a fair test. The identified levels of the new conceptualization of test fairness were entitled validity, construction and structure, administration, scoring, reporting, decision-making, consequences, security, explicitness, accountability, equality, and rights. The need for advancing context-specific and locally agreed-upon equity principles as driven by the impossibility of the fulfillment of the equality principle in the real world conditions is an important finding of this study with concrete implications for both theory and practice in the field.

Keywords: test fairness; model; reconceptualization; grounded theory; quality principle; equity principle

ARTICLE HISTORY

Received: 5 Mar. 2023 **Revised version received:** 19 May 2023

Accepted: 20 June 2023 **Available online:** 1 July 2023

* Corresponding author: English Department, Humanities Faculty, Bu-Ali Sina University, Hamedan, Iran
Email address: m.ahmadisafa@basu.ac.ir

10.30466/ijltr.2023.121333

Introduction

Testing has become an integral part of the life of the majority of people all over the world (McNamara et al., 2019). Such heightened stance and widespread use of testing and evaluation call for immediate attention to the qualities of a test, including its fairness (McNamara et al., 2019). Along with some allegedly philosophical underpinnings (e.g., Rawls, 1999, 2001), test fairness has its origin in Messick's (1989) validity as a unitary concept.

As a major concern for testing specialists in particular and educationalists in general, the definition of test fairness has evolved over time. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA et al., 2014) have provided an outline of the prevalent perspectives on fairness including (a) equitable treatment during the testing process, (b) lack of measurement bias, (c) access to the test construct(s), and (d) validity of individual test score interpretations for the intended uses. Besides, as stated in the *Standards for Quality and Fairness* (*Educational Testing Service [ETS]*, 2002, p. 17), fairness is conceptualized as "treating people with impartiality regardless of personal characteristics such as gender, race, ethnicity, or disability that are not relevant to their interaction with ETS." A more recent update of the standards (ETS, 2014, p. 19) mentions that a practically informative definition of fairness for test designers is "the extent to which the inferences made on the basis of test scores are valid for different groups of test takers." From another perspective, fairness is characterized as comparable validity for groups or individuals at every stage of testing and assessment (Xi, 2010).

Notwithstanding the evolutions of different definitions, owing to the tricky nature of the concept (Cole & Zieky, 2001; Zieky, 2006), there is still a lack of a single definitive technical and generally accepted conceptualization of test fairness. Besides, the debate has long prevailed as to the constituting factors of test fairness concept (e.g., Davies, 2010; Kane, 2010; Kunnan, 2008; Xi, 2010), with the types and nature of the constituting components of the concept being still somewhat elusive (Pommerich, 2016).

In the milieu of such uncertainty, tentative models of test fairness and their componential elements are proposed to maximally clarify it from given vantage points. As an eminent instance in this regard, the test fairness model of Kunnan (2000, 2004, 2008, and 2010) specifies a number of distinct qualities, including validity, absence of bias, access, administration, and social consequences as the componential constituents of the concept. Although this rather comprehensive model has been among the most frequently cited frameworks in the literature, the researchers' in-depth review of the related literature did not result in the identification of any empirical investigation into the credibility of this framework. Moreover, given the attested significance of attitudinal aspects of test fairness (Cole & Zieky, 2001), the literature documents a quite limited body of evidence on the attitude of the stakeholders towards the characteristics of a fair test.

On this basis and following a grounded theory approach, the present study examined different testing specialists and stakeholders' perspectives on the characteristics of a fair test. Moreover, looking at the concept from multiple vantage points, the study aimed to map the Test Fairness Framework (TFF) of Kunnan (2004, 2008, 2010) against the obtained empirical data in an attempt to revisit the validity of the model.

Theoretical background

Three Dominant Approaches to the Link between Validity and Test Fairness

Test fairness has stirred up a lot of controversy over the years and different interpretations of this concept have been proposed in educational assessment and measurement (e.g., AERA et al., 1999) and language testing contexts (Kunnan, 2000, 2004). On this basis, three approaches to test fairness have been dominant in educational assessment and testing literature: (1) fairness as an independent test feature, distinct from validity (AERA et al., 1985; Joint Committee on Testing Practice, 1988, 2004; Kane, 2010), (2) fairness as an all-encompassing test feature consisting of diverse aspects including validity (Kunnan, 2000, 2004, 2008), and (3) validity as the central test feature that fairness is straightly linked to (AERA et al., 1999, 2014; Davies, 2010; ETS, 2014, 2016, 2022; Messick, 1989; Willingham, 1999; Xi, 2010). An essential point on which the three conceptualizations of fairness fluctuate is how validity and fairness are related (Xi, 2010). In other words, these conceptualizations could be distinguished based on whether fairness is independent of validity, includes it, or is an aspect of it (Xi, 2010).

Kunnan's Test Fairness Framework (TFF)

Kunnan (1997, 2000, 2004, 2008, 2010), as a proponent of the second view, indicates that the validation frameworks are not able to cover all intricacies of test fairness. He contends that embedding fairness in validity damages both conceptions and can bring about confusion or diffusion in the focus of research agendas. Hence, Kunnan (2004) presented an ethics-inspired Test Fairness Framework (TFF) consisting of five main qualities of "validity," "absence of bias," "access," "administration," and "social consequences". In Kunnan's TFF (2004, pp. 37-39), the *validity* was further classified into different types, including "content representativeness or coverage evidence" (relabeled as "content representativeness and relevance" in Kunnan, 2008, p. 235), "construct or theory-based validity," "criterion-related validity," and "reliability". The *absence of bias* was explained by "offensive content or language," "unfair penalization based on test taker's background," and "disparate impact and standard-setting". However, in Kunnan (2008, p. 237), the absence of bias was relabeled as "content or language," "disparate impact," and "standard setting". Moreover, in Kunnan (2004), *access* was categorized into various types such as "educational access," "financial access," "geographical access," "personal access," and "conditions or equipment access". The *administration* component referred to "physical conditions" and "uniformity or consistency" (relabeled as "physical condition," "uniformity," and "test security" in Kunnan, 2008, p. 238). Lastly, the *social consequences* were explained by "washback" and "remedies".

Kunnan (2005, 2008, and 2010) expanded his first model of test fairness and suggested a Test Context Framework (TCF) after recognizing that the TFF had failed to consider the surrounding contexts of the test use. The new expanded framework (Kunnan, 2005, 2008) took into account the broader economic, political, social, educational, cultural, legal, and technological contexts of tests.

Research Methods Used in Test Fairness Research

In the language testing and assessment field, test fairness studies have focused on various aspects of the concept. Some projects aimed at the detection of differential item functioning (DIF) in English language tests (e.g., Ercikan & Oliveri, 2013; Ferne & Rupp, 2007; Geranpayeh & Kunnan, 2007; Young et al., 2013) or the effects of bias on the test results (Aryadoust, 2016; O'Loughlin, 2002). From a methodological perspective, some studies have applied CFA or SEM statistical techniques (e.g., In'nami & Koizumi, 2012; Yan et al., 2018; Yoo et al., 2018; Young et

al., 2008), and still another group of bias-related studies has relied on the score equity assessment (SEA) analysis introduced by Dorans (2006) to evaluate population invariance or the extent that test scores have the same meaning among the test-taker groups (Kim & Kolen, 2010; Yoo et al., 2018).

Besides previously mentioned studies, a recent though infrequent undertaking of the researchers has been an in-depth analysis of test fairness employing qualitative research methods. For instance, Pusawati (2014) explored the opinions of students from diverse backgrounds about their perceived degrees of fairness in the Test of English as a Foreign Language (TOEFL), the majority of whom expressed negative attitudes, using words such as unjust and unfair to describe their feelings. They believed that the topics in the TOEFL test were field-of-study specific and required background knowledge of the topic or familiarity with the culture of the US. The use of technology in the administration of the test, time constraints, and inauthentic speaking tasks were mentioned as other factors leading to unfairness.

As another instance, Moghadam and Nasirzadeh (2020) examined the fairness of an English reading comprehension test using a mixed research method and with heavy reliance on the test fairness framework of Kunnan (2004). The test was examined in terms of the constituting aspects of TFF by employing a semi-structured interview.

Lastly, to assist a better understanding of test fairness, Barrance (2019) investigated students' accounts of how fairness is enacted within their contexts, revealing the effect of several factors such as assessment environment on the assessments' fairness and the students' performance.

As a conclusive remark of this brief literature review and as McNamara et al. (2019) indicated, the number of tests that are designed considering test fairness principles seems to be quite limited, and the reasons behind this issue have been only scantly studied. This might be partly originating from the lack of an agreed-upon definition of test fairness, lack of certainty about the nature of test fairness, and lack of an approved model of test fairness (e.g., Davies, 2010; Kane, 2010; Kunnan, 2010; Xi, 2010). Against this backdrop, in the present study the researchers attempted to explore the empirical validity or the degree to which the accuracy of Kunnan's (2000, 2004, 2008) test fairness framework could be empirically corroborated through experimentation, accumulation of supporting research evidence, and systematic observation (in this case on the basis of the language testing specialists' and stakeholders' attitudes) rather than theory alone (VandenBos, 2015). They also aimed at revisiting different aspects and factors of the model on the basis of the empirical data obtained in the context of the study. For these purposes, the following research questions were formulated:

1. Is the test fairness framework of Kunnan (2008) empirically corroborated in typical Iranian educational contexts from different stakeholders' perspectives?

2. What is a valid model of test fairness in typical Iranian educational contexts?

Method

Participants

The participants of the study were10 university lecturers of TEFL, 20 high school English language teachers, 15 PhD students in TEFL, and 15 MA students in TEFL

who all participated in open-ended and semi-structured interviews. The rationale for the selection of the university TEFL lecturers and high school English language teachers was based on their previous experience in developing, administrating, scoring, reporting, and using different test scores during their career. The rationale for the selection of the higher education students of TEFL was based on their experience of taking different kinds of English tests during their studies. Besides, as the higher education students of TEFL in Iran pass courses on language testing and assessment, they were considered potentially eligible participants to discuss the subject of the study. The participants' age ranged from 25 to 55. The initial cohort of 60 participants comprised 34 females and 24 males. The interview participants were selected through convenience sampling from 35 different cities of around 24 provinces of Iran. The detailed demographic information of the participants is presented in Table 1.

Table 1
The Participants' Demographic Information

	Gender	Age Range	Frequencies	Total Frequencies/60
University TEFL Lecturers	Female	46-55	4	10
	Male	40-55	6	
High School Teachers	Female	25-50	11	20
	Male	35-53	9	
PhD Students	Female	28-40	10	15
	Male	29-37	5	
MA Students	Female	26-43	9	15
	Male	29-36	6	

Instruments and materials

Study data were collected using a series of researcher-made interviews. A mixture of open-ended and semi-structured interviews was carried out to delve deeply into participants' viewpoints on the needed qualities of a fair test. The interview questions were designed with an eye to the relevant literature in test fairness and the emergent ideas of the interviewees (Appendices A & B). Three TEFL experts reviewed and commented on the appropriateness of the interview questions, based on the ideas of whom needed modifications were made. Depending on the preference of the interviewees, English or Farsi was the medium of communication in the interviews. The interviews lasted for around 10 to 20 minutes for each participant.

Procedures

Prior to the start of the data collection phase, the researchers made any attempt to ensure ethical clearances, and obtain informed consent of all potential participants The participants received a brief explanation about the study and were notified that their participation was only on a voluntary basis. Moreover, they were assured of issues such as anonymity of the data, confidentiality of the results, and use of the data for research purposes only. The intended sample consisted of conveniently available Iranian stakeholders in the field of English language assessment and testing at different levels of expertise and experience.

Participation requests were sent to 3471 potential interview candidates through email and a variety of other social media applications (e.g., Telegram, Instagram, WhatsApp, LinkedIn, ResearchGate). An embedded URL link to the interview questions was forwarded to social media groups involving members such as university lecturers or experts, teachers, and higher education

students. Out of 3471 participation requests, 60 were affirmatively agreed. Of the whole range of responses, the data obtained from 43 respondents were gathered through online interviews, and the remaining respondents participated in telephone interviews.

The interview participants were requested to explain how they made sense of the test fairness. Concerning the analysis of the obtained data, as grounded theory was the first qualitative approach that had been procedurally rigorous and sufficiently elaborate enough to bear the criticisms of quantitative researchers, this methodologically elaborate approach was adopted as the analytical framework to provide an in-depth analysis of the phenomenon, generate inductive theoretical knowledge, make sense of test fairness complexity, and reach a fuller understanding of the concept (Dörnyei, 2007). Attempts were made to ensure grounded theory-based studies' criteria, including credibility, originality, resonance, and usefulness (Charmaz, 2006, 2014). Therefore, as for credibility, attempts were made to make sure that sufficient evidence was gathered in support of the claims. Regarding originality, the researchers made every attempt to assure that their theory could challenge, refine, and extend the existing knowledge, and concerning the resonance, they attempted to devise categories that offer deeper insights into the studied phenomenon. Besides, as for usefulness, they tried to make sure that the analyses could prompt further future research.

Considering the former criteria, the researchers transcribed, compared, and coded the responses (i.e., open coding, focused coding, axial coding, and selective coding). Reflection was made on what category each code belonged to or what category each piece of data signified. Due to the nature of the obtained data and the purposes of the research, an integration of different coding approaches, including a procedural approach (Strauss & Corbin, 1990) and a constructivist grounded theory method (Bryant, 2017; Charmaz, 2006, 2014) were employed, with heavy reliance on the constructivist assumptions and epistemological positions of Charmaz (2006, 2014) in approaching the research. Accordingly, some strategies such as initial/open coding, focused coding, axial coding, selective coding, or theoretical coding were used in the coding process. As the idea of coding in grounded theory differs from that of content analysis or quantitative text analysis in a number of important ways, one primary researcher/coder coded the data (under the supervision of the second researcher) and inter-coder reliability was not stimulated. Indeed, the calculation of the inter-coder reliability index was deemed inappropriate and technically impossible due to the epistemological background of grounded theory and the incremental, reflexive, and recursive nature of coding process in grounded theory (Oktay, 2012).

The memos or extended notes were written throughout the research process to help the researchers become more analytic and develop ideas by linking codes to conceptual categories. Then, the memos were integrated, prompting the researchers to move back and forth among data, analyses, and conceptualizations. The researchers found out that the obtained insights and the categories were engrossing but weak and imperfect. This finding provided stimulation to conduct theoretical sampling (Morse & Clark, 2019) to more adequately understand the circumstances in which test unfairness ensues. Once the obtained data saturated the categories, the researchers sorted the categories, followed leads, reached the fundamentals, and developed a theoretical rendering. During the whole process, the *"constant comparative methods"* of Glaser and Strauss (1967) were utilized to draw detailed comparisons at every phase of the analytic effort.

Results

The interview data were audio-recorded and transcribed verbatim to explore the validity of the TFF in Iranian educational context and to present a valid test fairness model on the basis of attitudinal empirical data in this context. Line-by-line, segment-by-segment, and incident-by-

incident initial coding processes were then used to produce a variety of information or thoughts and to come across ideas on which the researchers could construct their arguments. In the initial coding phase (also known as open coding), the basic meanings within every line or paragraph of the interview transcripts were determined by interpreting the reality embedded in the statements and actions of the participants. While remaining open to what is happening in the data and without applying preexisting concepts or categories in mind, each line or segment of data was named, creating codes of the significant data with simple, short, and precise words that described statements and opinions of the participants. In one of the interview transcripts, some initial codes stood out to the researchers. The related narrative and sample initial codes are displayed in Table 2.

Table 2
Initial/Open Coding

Initial/open Codes	Interview Data
Enough test-taking time, equal test-taking time, starting at a predetermined time, finishing at a predetermined time	Interviewer: What do you think the characteristics of the administration phase of a fair test are? Interviewee: Well, I think the time allocated to the exam should be sufficient so that the exam administrators do not take the students' exam papers while they are still responding to the questions. Besides, the test-taking time must be equal for all test takers. Sometimes, the proctors distribute the exam paper of some test takers late but collect their exam papers early, and you know, this is not fair at all. I myself had such a terrible experience, I received my exam paper ten minutes later than other test takers because the proctors had distributed the papers slowly, but in the end, one proctor speedily collected my paper as soon as the testing time finished. I suppose a fair test must also start and finish at a predetermined time.
Silence, lighting, quality of the chairs, distance of the chairs, quality of the amplifiers, air conditioning, degree of coldness, degree of warmth	Nothing should distract the test takers' attention. All proctors should sit down and not distract the test takers by walking or talking. Sometimes proctors speak close to the doors within the exam setting. Such behaviors would exert a negative influence on the performance of the test takers who sit in front of the door. The additional voices, such as transportation noise, etc., should not be heard in the exam setting. Totally, nothing should disturb the silence of the exam setting. Additionally, the exam setting conditions such as the light of the exam setting, quality of the chairs, distance of the chairs, quality of the amplifiers must be acceptable and equal for all the test takers. The air conditioning must be equal for all test takers. For example, The degree of coldness or warmth must be suitable and equal for all the test-takers in the exam rooms.

The initial codes extracted from the standpoint of each interviewee were compared with those of the next interviewees, and then focused coding was conducted, which is believed to be more abstract, selective, and directed compared with the initial coding phase (Glaser, 1978). In the focused coding phase, the most important or recurrent initial codes were used to construct the most outstanding categories while making the focused codes as concise and incisive as possible to

let them be considered as potential categories. The sample focused codes for the same excerpt previously narrated in Table 2 are presented in Table 3 below.

Table 3
Focused Coding

Focused Codes	Interview Data
Time	Interviewer: What do you think the characteristics of the administration phase of a fair test are? Interviewee: Well, I think the time allocated to the exam should be sufficient so that the exam administrators do not take the students' exam papers while they are still responding to the questions….
Physical setting and equipment	Nothing should distract the test takers' attention. All proctors should sit down and not distract the test takers by walking or talking. Sometimes proctors speak close to the doors within the exam setting. …

Next, the data that were coded in previous stages were reconsidered for axial coding purposes. Axial coding forms "a dense texture of relationships around the axis of a category" (Strauss, 1987, p. 64) and aims to categorize, integrate, arrange, and then bring the fractures of data back together afresh (Creswell, 1998). Indeed, initial coding disassembles data into separate codes, and axial coding reassembles data once more in a coherent whole (Charmaz, 2006; Strauss & Corbin, 1998). Thus, axial coding was carried out to link subcategories to categories (Strauss, 1987; Strauss & Corbin, 1990, 1998) and to transform texts into concepts that could be the basis for identifying larger category aspects (Charmaz, 2006; Strauss & Corbin, 1998). The subcategories subsumed under a category were determined and the relationships between them were displayed. A complete list of the axial codes is discussed in the following parts.

During the coding process, memo-writing was simultaneously carried out and aimed at uncovering unstated, implicit, or hidden meanings. For memo-writing purposes, the researchers stopped, analyzed, drew comparisons between data, codes, categories, and concepts, and wrote their ideas regarding the codes whenever they crossed their minds, arousing new and intriguing ideas that had not struck the researchers' minds earlier. The memos were critically reexamined and revised as the researchers proceeded, helping them to reconsider categories or find new ones in some cases.

Some questions concerning the codes emerged after several readings of the data, persuading the researchers to conduct emergent theoretical sampling and gather supplementary data with an intention to elaborate, explicate, strengthen, and improve the imperfect categories in the evolving theory. To this end, the researchers returned to research participants to ask further focused questions. Theoretical sampling was applied so long as no new properties appeared and the categories were saturated with new data. Afterward, in a back and forth movement process within the data, the researchers went back and recoded the same data, which in some cases generated new ideas. Such a back and forth process occurred several times due to the flexible nature of the grounded theory coding until the final codes, categories, and core concepts were discovered.

The following figures display the process through which open codes, focused codes, and axial codes were derived and related to each other. Figure 1 below displays the derived structure for the "validity" axial code.

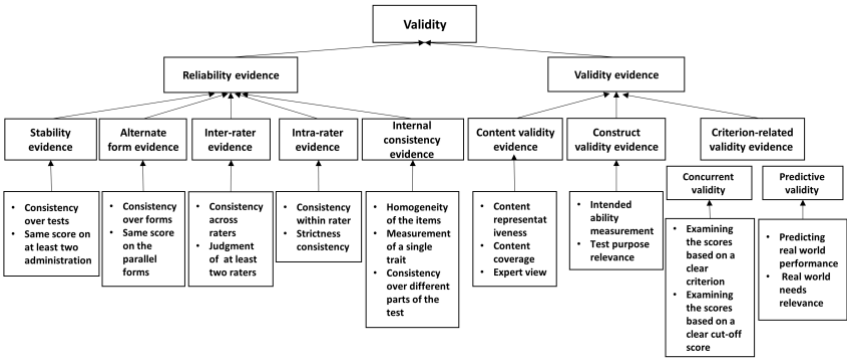

Figure 1. The Open Codes and Focused Codes Leading to the Validity Axial Code

As shown in Figure 1, a number of the most important initial codes were used to create the focused codes of "stability evidence," "alternate form evidence," "inter-rater evidence," "intra-rater evidence," "internal consistency evidence," "content validity evidence," "construct validity evidence," and "criterion-related validity evidence (subsuming lower-level focused codes of concurrent and predictive validity)." Then, more focused codes of "reliability evidence" and "validity evidence" were identified, on the basis of which the axial code of "validity" was selected to spot, integrate, and recognize the central themes in the statements, initial codes, and focused codes.

Figure 2 represents how open and focused codes led to "construction and structure" axial code.

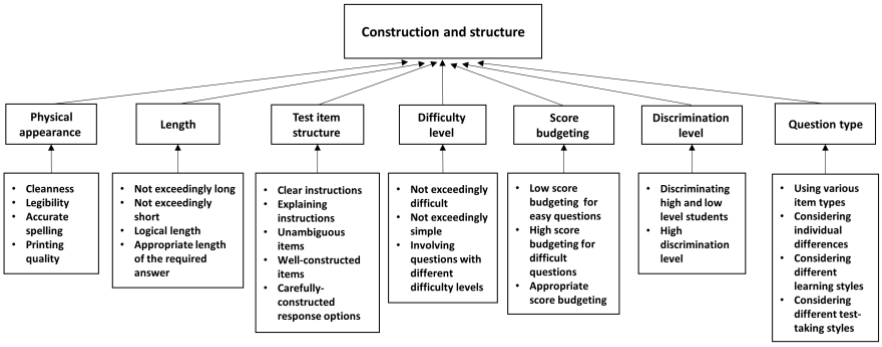

Figure 2. The Open Codes and Focused Codes Leading to the Construction and Structure Axial Code

As is demonstrated in Figure 2, "physical appearance," "length," "test item structure," "difficulty level," "score budgeting," "discrimination level" and "question type" were the focused codes to summarize the underlying initial/open codes. To unite these initial codes and the focused codes, the axial code of "construction and structure" was then derived.

Figure 3 displays how the open codes and focused codes fit together to form the "administration" axial code.

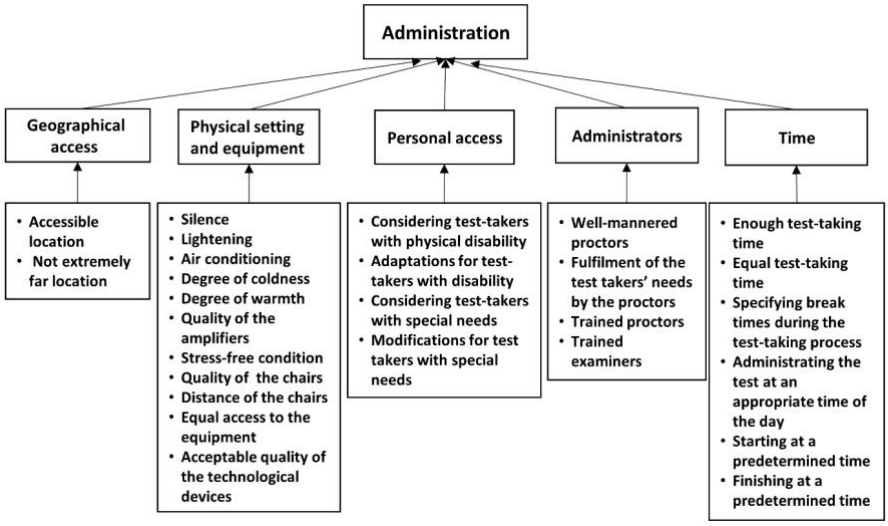

Figure 3. The Open Codes and Focused Codes Leading to the Administration Axial Code

As is evident above in Figure 3, one of the major categories or axial codes was "administration," the main properties of which include "geographical access," "physical setting and equipment," "personal access," "administrators," and "time."

Figure 4 displays open and focused codes that led to "scoring" axial code.

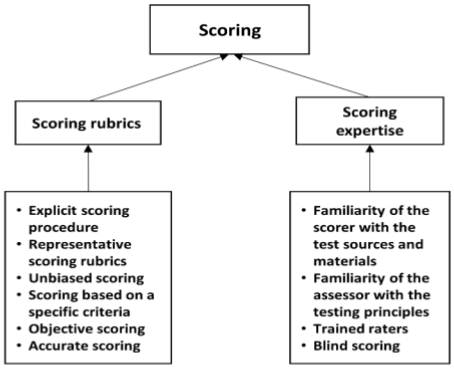

Figure 4. The Open Codes and Focused Codes Leading to the Scoring Axial Code

It is evident above in Figure 4 that the key category of "scoring" subsumed focused codes of "scoring rubrics" and "scoring expertise," and different initial/open codes as observable in the lowest rectangles in the Figure.

"Reporting" axial code and its related open and focused codes are displayed below in Figure 5.

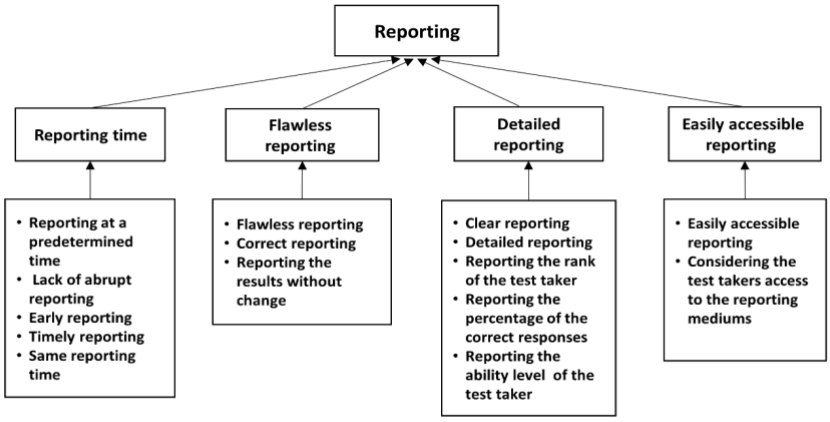

Figure 5. The Open Codes Leading to the Reporting Axial Code

As can be seen from Figure 5, the researchers advanced the codes of "reporting time," "flawless reporting," "detailed reporting," and "easily accessible reporting" under the major axial code of "reporting," with the relevant open codes shown in the lowest rectangles.

Figure 6 displays the constituting codes for the "decision-making" axial code.

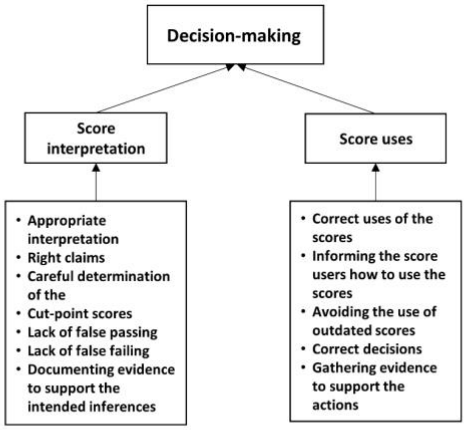

Figure 6. The Open Codes Leading to the Decision-making Axial Code

"Decision-making," was the axial code to embody "score interpretation" and "score uses," focused codes with the initial/open codes listed in the lowermost rectangles.

What follows in Figure 7 is the visual representation of the "consequences" axial code.

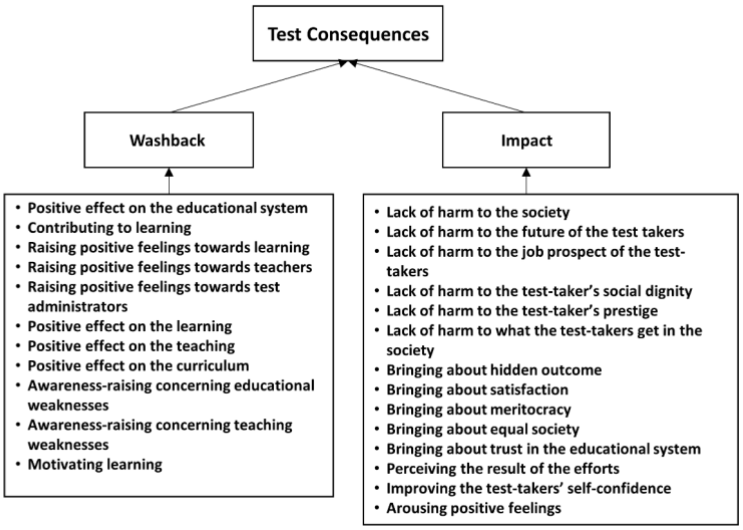

Figure 7. The Open Codes and Focused Codes Leading to the Test Consequences Axial Code

As is evident above, "washback" and "impact" were the major focused codes under "consequences" axial codes that were derived from rather long lists of open codes concerning the educational and social consequences of fair tests.

"Security" was the label of the next axial code derived from the analyses.

Figure 8. The Open Codes and Focused Codes Leading to the Security Axial Code

As displayed in Figure 8, "item security," "administration security," "scoring security," "reporting security," and "decision-making security" were the focused codes to summarize the underlying initial/open codes.

The structure of the "explicitness" axial code is displayed in Figure 9 below.

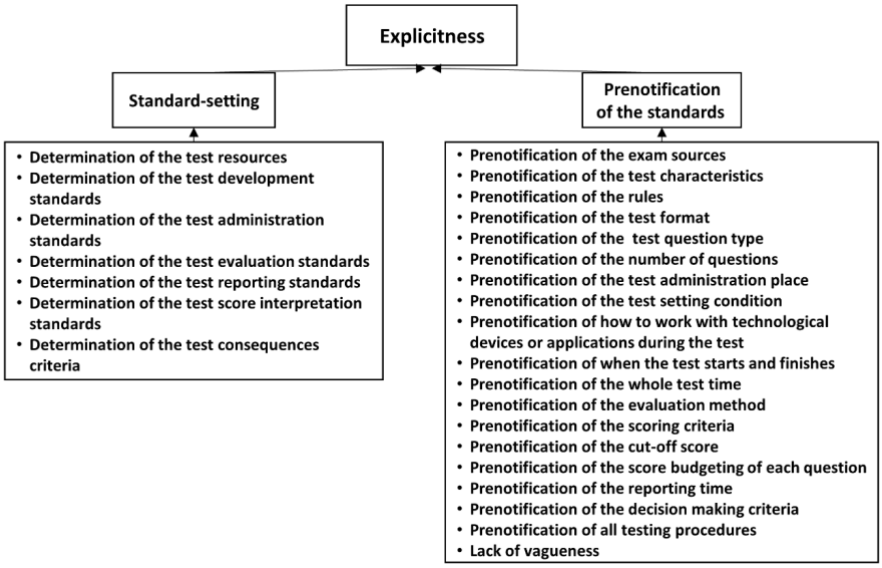

Figure 9. The Open Codes and Focused Codes Leading to the Explicitness Axial Code

As displayed in the Figure, "standard-setting" and "prenotification of the standards" were the two focused codes under the axial code of explicitness, which summarized the given underlying open codes. The next derived axial code was "accountability," which is displayed in Figure 10.

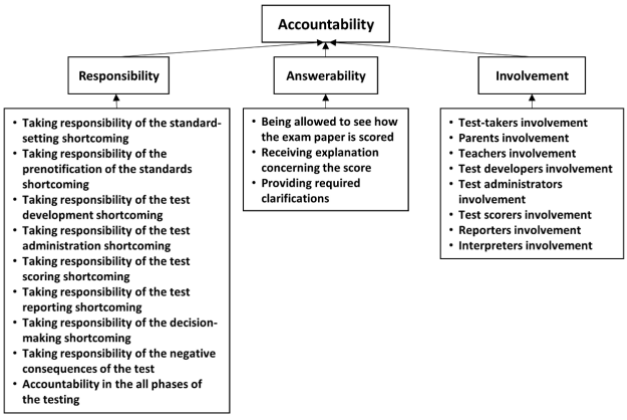

Figure 10. The Open Codes and Focused Codes Leading to the Accountability Axial Code

As demonstrated in Figure 10, while scrutinizing the accounts, "accountability" appeared as a common theme, forming a key category or axial code subsuming "responsibility," "answerability," and "involvement" as focused codes. Many initial/open codes were also discovered that are summarized in the lowermost level of Figure.

"Equality" was the next axial code derived from the data, the componential structure of which is evident below in Figure 11.

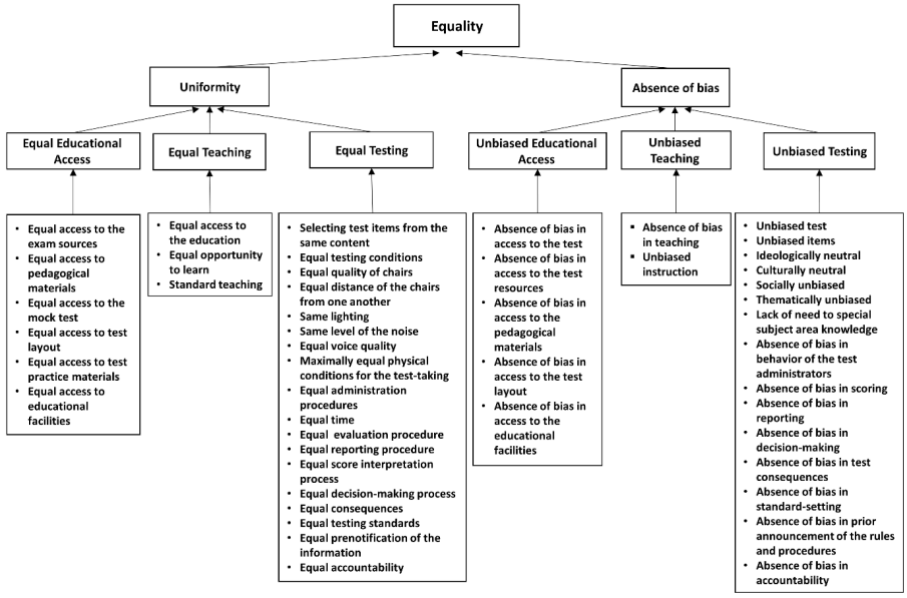

Figure 11. The Open Codes and Focused Codes Leading to the Equality Axial Code

The significance of equal and unbiased education in the fairness of a test was strongly underscored in the interview data. Therefore, to account for the diverse perspectives, a distinction was drawn between "education" and "testing." A number of the most significant initial codes were used to form the focused codes of "equal educational access," "equal teaching," and "equal testing" which suggested building a more focused code of "uniformity." Similarly, several other important initial codes were used to advance focused codes of "unbiased educational access," "unbiased teaching," and "unbiased testing" that were then united under the more focused code of "absence of bias." Ultimately, all the focused codes were assembled under the axial code of "equality."

The final axial code derived from the altitudinal data of the study to be incorporated under the test fairness construct was the rights of the test takers. Figure 12 represents how different codes came together under the axial code of "rights."

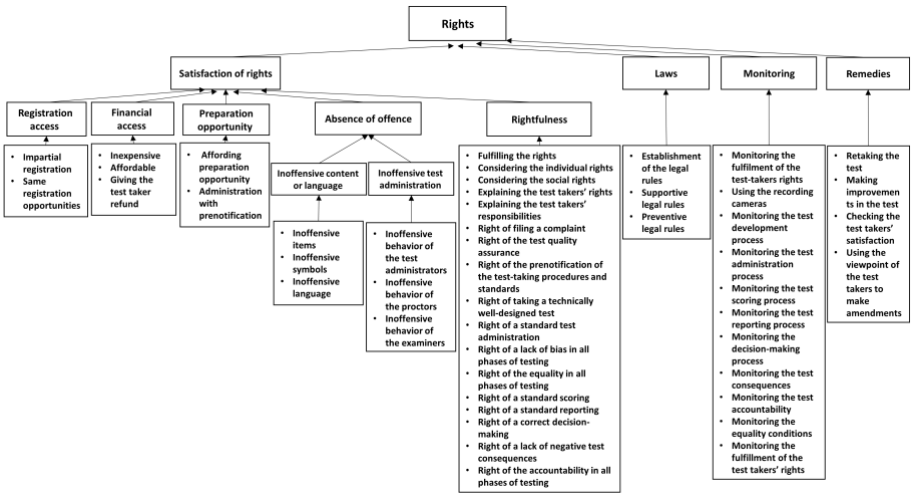

Figure 12. The Open Codes and Focused Codes Leading to the Rights Axial Code

As can be seen in Figure 12, the "rights" appeared as a category or an axial code, subsuming numerous open codes, first-order focused codes (i.e., "inoffensive content or language" and "inoffensive test administrators"), second-order focused codes (i.e., "registration opportunity," "financial access," "preparation opportunity," "absence of offense," and "satisfaction of rights"), and third-order focused codes (i.e., "rightfulness," "laws," "monitoring," and "remedies").

Ultimately, Figure 13 presents a visual representation of how all previously discovered axial codes united under the selective code of "fairness."

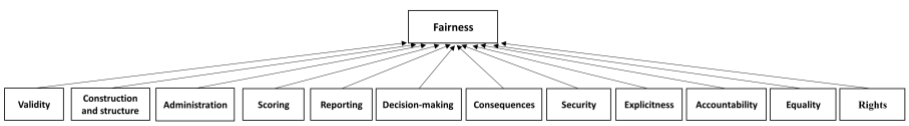

Figure 13. The Axial Codes Leading to the Fairness Selective Code

As is evident in the Figure above, the axial codes of "construction and structure," "administration," "scoring," "reporting," "decision-making," "consequences," "security," "explicitness," "accountability," "equality," and "rights" came together beneath the topmost selective category of "fairness." Table 4 below enlists all the main derived properties of the test fairness in more detail.

Table 4
The Main Properties of the Test Fairness

1. Validity 1.1. Reliability evidence Stability evidence Alternate form evidence Inter-rater evidence Intra-rater evidence Internal consistency evidence 1.2. Validity evidence Content validity evidence Construct validity evidence Criterion-related validity evidence Concurrent validity Predictive validity	2. Construction and structure Physical appearance of the test Length Test item structure Difficulty level Score budgeting Discrimination level Question type	3. Administration Geographical access Physical setting and equipment Personal access Administrators Time	4. Scoring Scoring rubrics Scoring expertise	5. Reporting Reporting time Flawless reporting Detailed reporting Easily accessible reporting	6. Decision-making Score interpretation Score uses
7. Consequences Washback Impact	8. Security Item security Administration security Scoring security Reporting security Decision-making security	9. Explicitness Standard-setting Prenotification of the standards	10. Accountability Responsibility Answerability Involvement	11. Equality 11.1. Uniformity Equal educational access Equal teaching Equal testing 11.2. Absence of bias Unbiased educational access Unbiased teaching Unbiased testing	12. Rights 12.1. Rightfulness Registration opportunity Financial access Preparation opportunity Absence of offense Inoffensive content or language Inoffensive test administration Satisfaction of rights 12.2. Laws 12.3. Monitoring 12.4. Remedies

The saturated categories were sorted and diagrammed to consolidate the evolving theory after digging out the codes and categories (Glaser, 1978; Glaser & Strauss, 1967; Strauss, 1987). To visually represent through flexible maps how the tentative relationships were organized (Konecki, 2019), the clustering technique was applied; thereby, the researchers liberated their creativity (Charmaz, 2006), wrote a key category or concept, and then circled and linked the key category to smaller circles to demonstrate the pertinent properties, associations, and weights (Rico, 1983). On this basis, Figure 14 below sketches the properties of a fair test and how they interact or overlap, identifying several categories under a single selective category of test fairness.

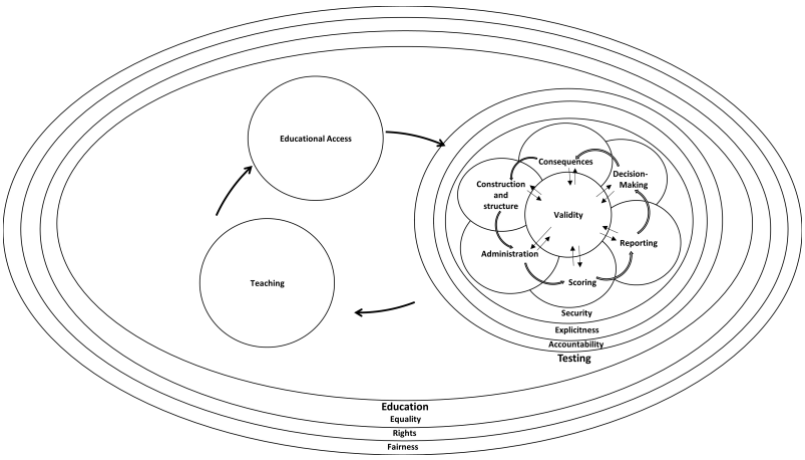

Figure 14. The Revised Test Fairness Model (RTFM)

As shown in Figure 14, numerous interactions arise at various levels. Equal teaching and equal educational access for all the test takers, being mentioned as essential prerequisites for test fairness by the majority of the interview participants, were added to the emerging theoretical model. This theoretical model takes into consideration the links and hierarchical order of the "validity," "construction and structure," "administration," "scoring," "reporting," "decision-making," "consequences," "security," "explicitness," "accountability," "equality," and "rights" for constructing a possible explanation of test fairness.

Discussion

This study set out to examine the validity of the TFF in the Iranian context and develop a revisited explanatory theoretical model of test fairness, thereby reconstructing conceptual understandings of the phenomenon. To this end, interviews were held with different stakeholders and the obtained attitudinal data were scrutinized, and maximally reflecting codes were assigned at different coding levels. The major categories were constructed after coding, memoing, and forming the elementary categories and then they were raised to theoretical concepts through more analytic refinement. In this process, obscurities, gaps, questions, and new ideas constantly emerged, necessitating theoretical sampling for clarification of the esoteric areas. The researchers alternated successively between data collection and data analysis and probed if there were any remarks in the data that did not fit the analysis or model, focusing on thought-provoking deviant responses so as to take different vantage points into consideration and present alternative justifications or refinements in the emerging theory.

A relative one-to-one correspondence was quite evident between the perspectives of the study participants and the TFF of Kunnan (2008); however, the existing categories of Kunnan (2008) could not justify the whole range of relevant personal experiences or accounts of the participants. For this reason, the researchers concluded that TFF was in need of expansion to a broader and more inclusive model by adding, replacing, and relabeling some of the concepts.

The newly formulated model or the Revised Test Fairness Model (RTFM) identified "validity," "construction and structure," "administration," "scoring," "reporting," "decision-making," "consequences," "security," "explicitness," "accountability," "equality," and "rights" as its componential factors, and a hierarchy of links was postulated between the properties, accommodating the core categories or the theoretical concepts that came under the test fairness umbrella term.

Some of the core categories and the subcategories of the derived revisited model of test fairness, which were either new additions or restructurings of the components of TFF, are further discussed below. Kunnan (2008, 2010) used the reliability evidence term to refer to "internal consistency," "inter-rater," "alternate forms," and "stability." Nevertheless, based on the accounts of the participants of the present study and in line with Bachman (1990, p. 178) and Farhady et al. (2012, p. 134), "intra-rater evidence" needs to be added to the model. The notion of "validity evidence" remains constant with no difference from Kunnan's TFF (2004, 2008).

Moreover, according to the interview findings and consistent with Farhady et al. (2012, pp. 126-141), the subcategory of "construction and structure," which was not originally discussed in TFF is included in RTFM.

The subcategories of "access" in Kunnan's TFF (2008), including "geographical," "personal," and "equipment and conditions," are transferred to "administration" in the revised model, with "equipment and conditions," being integrated to the "physical setting" forming the category of

"physical setting and equipment." It is noteworthy that as far as "personal access" is concerned, a systematic description of the notion seems to be still lacking. For Kunnan (2008), "personal access" refers to whether necessary accommodations are prepared for the test takers with physical and learning disabilities, thereby directing exclusive attention to the needs of the individuals with disabilities and neglecting the test-takers with special needs such as the left-handed test takers who are in real need of left-handed armchairs during the test-taking process. This finding broadly supports the studies which called for modifications in test construction, design, and administration to accommodate test takers with special needs (e.g., Armstrong, 2018; Butler & Iino, 2021). Furthermore, "uniformity" category is transferred from the "administration" in TFF to the newly developed category of "equality" in the revised model. The reason is that "uniformity" seemed to be a higher-order concept that was possibly necessary not only in the administration process of a test but also in other stages of the testing process, including the test development, test scoring, reporting of the test results, test-based decision-making, test-related consequences, explicit standards setting, prenotification of the standards, and accountability. This reflects the stance of Menken (2017) and Leung and Lewkowicz (2017) who underscored the importance of uniformity of the testing experience for all test takers, with increased attention to the uniformity of approach and content and uniformity of rating. Besides, "security" is relocated to a higher level in RTFM. To justify the decision, it is necessary to state that Kunnan (2008) merely mentioned the insecurity of test administration or test contents as an issue of test security. Our point, however, similar to Wollack and Case (2016), is that besides pre-administration and while-administration security, the post-administration security of the test results, test-based decisions, and personal information of the test takers are possibly other important aspects that deserve due consideration. It is worth mentioning that the "administrators" and "time" are also additional new subcategories in RTFM model, the importance of which is supported by AERA et al. (2014, p. 65) and Farhady et al. (2012, p. 145).

A number of categories included in RTFM were already included in earlier conceptualizations of test fairness. Such categories are "scoring" (ETS, 2014), "reporting" (International Test Commission, 2014; Wollack & Case, 2016), and "decision-making" (Liu & Dorans, 2016) which were not included in Kunnan's TFF (2008) framework. Concerning "scoring," the attitudinal data obtained underscored that scale points needed to be so specified that distinct levels be clearly discernable while scoring, and the raters needed to be educated and develop the expertise to distinguish among the different levels. Moreover, a comparison of the findings with those of the literature (e.g., AERA et al., 2014; ETS, 2014) confirmed the significance of detailed reporting and timely access to score reports.

Kunnan (2008) classified "washback" and "remedies" under the heading of the consequences. From Kunnan's perspective (2008, p. 238), washback refers to "the effect of a test on instructional practices, such as teaching, materials, learning, test-taking strategies, etc.". From this point of view, washback mainly focuses on the effect of test on teaching and learning (Andrews, 2004; Bachman, 1990) and disregards the social dimensions of test uses and consequences (Shohamy, 2017); however, on the basis of our findings and consistent with some recent works, the effects of a test are not confined only to teaching and learning (Tsagari & Cheng, 2017). Tests might bring about many wider impacts and consequences for the test takers, including social, psychological, economic, etc. (Shohamy, 2017; Tsagari & Cheng, 2017). As a result, another label named test "impact" is added to "consequences." Moreover, in the RTFM, "remedies" is relocated under a higher-order category called "rights." This relocation partly echoes Phillips' (2016) perspective that "remedies" is an entirely relevant ethical and legal aspect of fair testing.

In view of "standard-setting," Kunnan believed that "test developers and score users need to be confident that appropriate measures and statistically sound and unbiased selection models are in use" (2008, p. 237). Comparison of this perspective with those of others (e.g., AERA et al., 2014) confirms that this definition of the standard-setting is confined only to the test development and

decision-making process; however, similar to the perspectives of the International Test Commission (2014) and Wollack *and Case (*2016), our findings indicate that accuracy and standard-setting are not limited only to the test development process and selection decisions, rather they are crucial in all stages of testing, involving the test development, test administration, test scoring, reporting of the test results, test-based decision-making, and test-related consequences. On this basis, the "standard-setting," which was classified under the heading of "absence of bias" in Kunnan's (2008) framework, is relocated under the category of "explicitness" in the RTFM.

Consistent with AERA et al. (2014), the current study confirmed that the standards need to be appropriately determined and announced to all stakeholders in all stages of the testing process. The Standards (AERA et al., 2014) are clear that "[t]hose responsible for testing should adhere to standardized test administration, scoring, and security protocols so that test scores will reflect the construct(s) being assessed and will not be unduly influenced by idiosyncrasies in the testing process" (p. 65). In the same vein, Bachman (1990) contends that the processes or conditions followed in eliciting and observing performance must be determined precisely alongside standard procedures for administering and scoring the test.

One of the beneficial impacts of standard-setting to test fairness is that explicit standards "are likely to result in increased understanding by and trust on the part of the public" (Cizek & Bunch, 2007, p. 8). It is also helpful to test fairness in that standard-setting intends to make sure that all the stakeholders meet the standards and scorers are not influenced by factors extraneous to the response (Brown, 1996). In other words, standard-setting attempts to prevent multiple interpretations and tries to ensure that test results and decision-making processes are not dependent on the personal ideas of the stakeholders (Brooks, 2017; Konrad et al., 2018).

On the other hand, as the exact link between TFF and Test Context Framework (TCF) was not visually represented in Kunnan (2008), to partly represent the linkage between the main qualities of these two frameworks or the relationship between the testing process and the socio-political context, "accountability" is embedded in RTFM as a higher-order concept embodying several lower-order ones. In line with AERA et al. (2014) and Hamilton and Koretz (2002), this study considers a test-based accountability system as an indispensable mechanism for the sustained efficiency of each educational program.

Moreover, according to the findings of the present study, it seemed that Kunnan (2008) failed to consider the "absence of bias" and "equality" under a broad inclusive term. Accordingly, the category of "absence of bias" in Kunnan (2008) is transferred under one of the subcategories of the higher-order category of "equality" in the RTFM. The reason behind the transposition is that on the basis of the interview findings, it was assumed that bias could not be a narrow concept; that is, lack of bias seemed to be necessary not only in the test development processes but also in terms of administration, scoring, reporting, decision-making, accompanied consequences, security, explicitness, and accountability. This finding is consistent with those of Van de Vijver and Poortinga (2005), Van de Vijver and Tanzer (2004), and Wollack *and Case (*2016), who mention administration and rater bias as other related sources of bias associated with specific groups besides construct and item bias. Additionally, according to the interview findings, equal learning opportunities seemed to be highly important in a theory of test fairness. It is worth mentioning that variations in viewpoints concerning equality were noticed once the researchers compared the perspectives of different interview participants. All of the participants embraced equality as a prerequisite for test fairness. However, some of the participants shared an implicit agreement that equality, and particularly equal learning, was an ideal assumption for the fulfillment of the test fairness that seemed not to be completely achievable in the real world. This finding is corroborated by Rawls (1999), who believes that real-world opportunities are unequal for different people. Furthermore, on the basis of the interview data, "educational access" was

transferred from the category of "access" in the framework of Kunnan (2008) to the category of "equality" in the current model.

According to the research findings, we suppose that one golden rule of test fairness seems to be the equality principle. Test fairness will not probably be attainable if equality is deviated. Thus, creating equal conditions for all test-takers in all testing phases including before, whilst, and after test administration seems to be necessary. However, similar to Davies (2010), we think the realization of equality and fairness is not simply and perfectly achievable in the real world, necessitating the need for the equity principle. This view seems consonant with Rawls (1999), whose main focus was on the significance of fair opportunity, permitting inequality in rewards if the inequalities eventually remediate the disadvantaged members. In line with Chapelle (2021) and Inbar-Lourie (2017), we suppose locally agreed-upon equity principles need to be developed according to the context-specific needs (Sen, 1992) of the stakeholders and based on the situational needs of each context.

As Spolsky (1981) contends, ethics and rights of the test takers are essential themes in language assessment and testing and, on the basis of this study's findings, are strongly in need of inclusion in the test fairness models. Consistently, Davies (2017, p. 411) maintains that the test takers need to be entitled "the right of access to the legal system for protection and claims against others, for defense against charges, for protection of the law, and for equality of treatment under the law". Rooted in such theoretical predisposition and on the basis of the obtained empirical data, categories of "rightfulness," "laws," "monitoring," and "remedies" are included in the RTFM, with the "financial access" and "offensive content or language" in Kunnan's (2008) TFF being relocated and reclassified under the "rightfulness". For Kunnan (2008), the language, dialect, or content must not be biased or offensive against test takers from different backgrounds.

Furthermore, addressing test-takers' rights, the "registration opportunity," "preparation opportunity," and "satisfaction of rights" are added to the model on the grounds that based on the interviewees' accounts and in accordance with Stone and Cook (2016), if the rights of the test takers are to be fulfilled, all test takers need to be given the opportunity to register for the test and the opportunity to prepare for the test. In addition, and on the basis of study findings and consistent with the perspective adopted by Bachman (1990), issues related to test ethics were found context-and-culture-dependent and very complex. Overall, these findings entail a number of theoretical and pedagogical implications that are discussed briefly in the following part.

Conclusion

In conclusion, empirical evidence of the study heightened the need for inclusion of what has been termed "construction and structure," "scoring," "reporting," "decision-making," "consequences," "security," "explicitness," "accountability," "equality," and "rights" in the investigation of test fairness as it seems that a framework overlooking these properties might fail to justify the whole range of experiences with unfairness. Moreover, consistent with Phillips (2016), the study underscores the need for remediation opportunities for violation of rights (e.g., absence of instruction on the tested skills, inadequate advance notice, etc.).

In general, our findings confirm that the focus on the test itself alone is not probably sufficient for the development of a fair test and a theory of test fairness needs to make reference to the influence of the wider contextual factors if it is to be a comprehensive and valid one. The revisited theoretical model of test fairness highlights a number of implications in the field of educational testing and assessment in general and language testing in particular among which the following might be the most salient:

- In a theory of test fairness, the test construction and structure need to be considered.
- For the administration of a fair test, time budgeting is a crucial factor.
- The scoring of a fair test needs to be done by trained raters and on the basis of objective and appropriately determined criteria.
- A fair test reports the results flawlessly and in detail at a predetermined time through a medium accessible to the test takers.
- In a fair test, the applications or uses of the test scores are clearly determined to the extent possible.
- A fair test observes not only pre-administration security but also while-administration and post-administration security.
- A fair test appropriately determines and prenotifies the standards pertaining to all phases of the testing process (e.g., test development, administration, scoring, reporting, decision-making, consequences, and security).
- In a fair test, test-giving authorities are accountable in and for all phases of the testing process.
- On the basis of the study findings, the development and administration of a fair test are not possible without educational fairness. The classroom content and teaching process need to be optimally equal for all test takers; otherwise, the test could be hardly considered fair as unequal education may have serious negative consequences for language testing.
- A fair test observes the rights of the test takers in all testing phases.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. J. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-53). Lawrence Erlbaum Associates, Inc.

Armstrong, D. (2018). Am I just stupid? Key issues for teachers involved in high-stakes testing with children who have dyslexia. In D. Xerri & P.V. Briffa (Eds.), *Teacher involvement in high stakes language testing* (pp. 67-82). Springer. https://doi.org/10.1007/978-3-319-77177-9_5

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly, 13*(1), 1–24. https://doi.org/10.1080/15434303.2015.1133626

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barrance, R. (2019). The fairness of internal assessment in the GCSE: The value of students' accounts. *Assessment in Education: Principles, Policy & Practice, 26*(5), 563-583. https://doi.org/10.1080/0969594X.2019.1619514

Brooks, R. L. (2017). Language assessment in the US government. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 64-76). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_4

Brown, J.D. (1996). *Testing in language programs*. Prentice Hall.

Bryant, A. (2017). *Grounded theory and grounded theorizing: Pragmatism in research practice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199922604.001.0001

Butler, Y.G., & Iino, M. (2021). Fairness in college entrance exams in Japan and the planned use of external tests in English. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in language testing around the world: Insights for language test users* (pp. 47-56). Springer. https://doi.org/10.1007/978-981-33-4232-3_5

Chapelle, C. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11-20). Taylor & Francis. https://doi.org/10.4324/9781351034784-3

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE Publications, Inc.

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985918

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, *38(4),* 369–382. https://doi.org/10.1111/j.1745-3984.2001.tb01132.x

Davies, A. (2010). Test fairness: A response. *Language Testing, 27*(2), 171–176. https://doi.org/10.1177/0265532209349466

Davies, A. (2017). Ethics, professionalism, rights, and codes. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 397-415). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_27

Dorans, N. J. (2006). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, *41*(1), 43–68. https://doi.org/10.1111/j.1745-3984.2004.tb01158.x

Dörnyei, Z. (2007). *Research methods in Applied linguistics*. Oxford University Press.

Educational Testing Service. (2002). *ETS standards for quality and fairness*.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. https://www.ets.org/s/about/pdf/standards.pdf

Educational Testing Service. (2016). *ETS international principles for the fairness of assessments: A manual for developing locally appropriate fairness guidelines for various countries*.

Educational Testing Service. (2022). *ETS guidelines for developing fair tests and communications*.

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 69–86). Emerald Publishing.

Farhady, H., Jafarpur, A., & Birjandi, P. (2012). *Testing language skills from theory to practice*. The Organization for Researching and Composing University textbooks in the Humanities.

Ferne, T., & Rupp, A. (2007). A synthesis of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113–148. https://doi.org/10.1080/15434300701375923

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

Glaser, B. (1978). *Theoretical Sensitivity*: *Advances in the methodology of grounded theory*. Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Aldine Transaction. https://doi.org/10.1097/00006199-196807000-00014

Hamilton, L. S. & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp.13-49). RAND Corporation.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 257–270). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_19

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC test: A multiple sample analysis. *Language Testing*, *29*, 131–152. https://doi.org/10.1177/0265532211413444

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195–217. https://doi.org/10.1080/15305058.2014.918040

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* https://www.apa.org/science/programs/testing/fair-testing.pdf

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kim, H. K., & Kolen, M. J. (2010). The effect of repeaters on equating. *Applied Measurement in Education*, *23*, 242-265. https://doi.org/10.1080/08957347.2010.486024

Konecki, K.T. (2019). Visual images and grounded theory methodology. In A. Bryant & K. Charmaz (Eds.), *The SAGE handbook of current developments in grounded theory* (pp. 352-373). SAGE Publications Ltd. https://doi.org/10.4135/9781526485656.n19

Konrad, E., Spöttl, C., Holzknecht, F., & Kremmel, B. (2018). The role of classroom teachers in standard setting and benchmarking. In D. Xerri & P.V. Briffa (Eds.), *Teacher involvement in high stakes language Testing* (pp. 11-29). Springer. https://doi.org/10.1007/978-3-319-77177-9_2

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta (Ed.), C*urrent developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 85–105). University of Jyväskylä.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge University Press.

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

Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 779–794). Routledge.

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

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

Leung, C., & Lewkowicz, J. (2017). Assessing second/additional language of diverse populations. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 343-358). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_23

Liu, J., & Dorans, N. J. (2016). Philosophical perspectives on fairness in educational assessment. In J. D. Neil & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp.97-108). Routledge.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.

Menken, K. (2017). High-stakes tests as de facto language education policies. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 343-358). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_25

Messick, S. A. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp.13- 103). Macmillan.

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's Test Fairness Framework (TFF) on a reading comprehension test. *Language Testing in Asia, 10*(7), 1-21. https://doi.org/10.1186/s40468-020-00105-2

Morse, J. M., & Clark, L. (2019). The nuances of grounded theory sampling and the pivotal role of theoretical sampling. In A. Bryant & K. Charmaz (Eds.), *The SAGE handbook of current developments in grounded theory* (pp. 145-166). SAGE Publications Ltd. https://doi.org/10.4135/9781526485656.n9

Oktay, J. S. (2012). *Grounded theory*. Oxford University Press.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*)2(, 169–192. https://doi.org/10.1191/0265532202lt226oa

Phillips, S. E. (2016). Legal aspects of test fairness. In J. D. Neil & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 239-266). Routledge. https://doi.org/10.4324/9781315774527-16

Pommerich, M. (2016). The fairness of comparing test scores across different tests or modes of administration. In J. D. Neil & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 111-134). Routledge. https://doi.org/10.4324/9781315774527-9

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

Rawls, J. (1999). *A theory of justice (Rev. ed.)*. Harvard University Press. https://doi.org/10.4159/9780674042582

Rawls, J. (2001). *Justice as Fairness: A restatement*. Harvard University Press. https://doi.org/10.2307/j.ctv31xf5v0

Rico, G. L. (1983). *Writing the natural way: Using right-brain techniques to release your expressive powers*. New York: St. Martin's Press.

Sen, A.(1992). *Inequality reexamined*. Harvard University Press.

Shohamy, E. (2017). Critical language testing. In E. Shohamy, L. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 441–454). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_26

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

Stone, E. A., & Cook, L. (2016). Testing individuals in special populations. In J. D. Neil & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp.157-181). Routledge. https://doi.org/10.4324/9781315774527-11

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge University Press. https://doi.org/10.1017/CBO9780511557842

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). SAGE Publications, Inc.

Strauss, A. L. & Corbin, J. (1998). *Basics of qualitative research: Grounded theory procedures and techniques* (2nd ed.). SAGE.

Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 359-372). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_24

VandenBos, G. R. (2015). *APA dictionary of psychology* (2nd Ed.) American Psychological Association.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Lawrence Erlbaum Associates.

Willingham, W. W. (1999). A systemic view of test fairness. In S. J. Messick (Ed.), *Assessment in higher education: Issues in access, quality, student development, and public policy* (pp. 213–242). Lawrence Erlbaum Associates.

Wollack*, J. A., & Case,* S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 33-53). Routledge. https://doi.org/10.4324/9781315774527-4

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

Yan, X., Cheng, L., & Ginther, A. (2018). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing, 36*(2). 207-234. https://doi.org/10.1177/0265532218775764

Yoo, H., Manna, V. F., Monfis, L. F., & Oh, H. J. (2018). Measuring English language proficiency across subgroups: Using score equity assessment to evaluate test fairness. *Language Testing, 36*(2), 289-309. https://doi.org/10.1177/0265532218776040

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13(2-3)*, 170–192. https://doi.org/10.1080/10627190802394388

Young, J. W., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). *Assessing the test information function and differential item functioning for TOEFL Junior Standard* (TOEFL Junior® Research Report TOEFL JR–0). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02324.x

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Lawrence Erlbaum Associates.

Appendix A: The interview questions

1. Have you ever felt that a test has been unfair (to you or others)? Why? Please explain.

2. What do you think the characteristics of a fair test are?

3. What do you think the characteristics of the preparation phase of a fair test are?

4. Do you think equal learning is a prerequisite for a fair test? Please explain.

5. What do you think the characteristics of the content of a fair are?

6. What do you think the characteristics of the administration phase of a fair test are?

7. What do you think the characteristics of the scoring phase of a fair test are?

8. What do you think about score adjustment for minority groups? Is it fair?

9. What do you think the characteristics of the reporting phase of a fair test are?

10. What do you think the characteristics of the decision-making phase of a fair test are?

11. How do you think the consequences of a fair test should be?

12. Do you see any relationship between test administrators' (e.g., test developers, teachers,…) accountability and test fairness? Please explain.

13. Do you think the test takers' rights need to be fulfilled in a fair test? How?

14. Do you have any suggestion or comment concerning how to improve fairness of a test?

Appendix B: Persian version of the interview questions

سوالات مصاحبه

1. تا کنون احساس کرده‌اید که آزمونی (نسبت به خود یا دیگران) ناعادلانه بوده است؟ چرا؟ لطفا توضیح دهید.

2. به نظر شما ویژگی‌های یک آزمون عادلانه چه هستند؟

3. به نظر شما ویژگی‌های مرحله آمادگی برای آزمون عادلانه چه هستند؟

4. به نظر شما یادگیری برابر پیش‌نیاز آزمون عادلانه است؟ لطفا توضیح دهید.

5. به نظر شما ویژگی‌های محتوای آزمون عادلانه (سوالات آزمون) چه هستند؟

6. به نظر شما ویژگی‌های مرحله اجرای آزمون عادلانه (شرایط امتحان، آزمون‌گیران، مراقبین و ...) چه هستند؟

7. به نظر شما ویژگی‌های مرحله نمره‌دهی آزمون عادلانه چه هستند؟

8. به نظر شما تعدیل نمرات (برای گروه‌های اقلیت) در آزمون‌های عادلانه کار درستی است؟ چرا؟

9. به نظر شما ویژگی‌های مرحله گزارش نتایج آزمون عادلانه چه هستند؟

10. به نظر شما ویژگی‌های مرحله تصمیم‌گیری (مثلا قبولی یا مردودی) آزمون عادلانه چه هستند؟

11. به نظر شما پیامدهای یک آزمون عادلانه چگونه می‌باشد؟

12. از نظر شما آیا میان مسئولیت‌پذیری و پاسخگو بودن برگزارکنندگان آزمون (طراحان آزمون، معلمین و ....) و عادلانه بودن آزمون رابطه ای وجود دارد؟ لطفا توضیح دهید

13. آیا به نظر شما حقوق آزمون دهندگان در یک آزمون عادلانه مورد توجه قرار میگیرد؟ چگونه ؟

14. آیا پیشنهادی برای عادلانه‌تر کردن آزمون ها دارید؟

Shima Beheshti is PhD candidate of TEFL at Bu-Ali Sina University. As an enthusiastic researcher and a university lecturer, she is interested in second language assessment, specifically test fairness, dynamic assessment, and the interconnection between test performance and collaborative dialogue.

Mohammad Ahmadi Safa is associate professor of Teaching English as a Foreign Language at Bu-Ali Sina University of Hamedan. He has widely published articles on different aspects of foreign language pedagogy including sociocultural orientations in language education and evaluation, Interlanguage Pragmatics, and Intercultural Communicative Competence development in foreign language education contexts.