Fairness and Bias in Language Testing

Malihe Yaghoobi
Department of ELT, Ayatollah Amoli Branch, Islamic Azad University, Amol, Iran
Email: Malihe_Yaghoobi21@yahoo.com
*Corresponding author’s email: Malihe_Yaghoobi21@yahoo.com

Abstract
In the process of language teaching and learning, testing plays a crucial role. A good test should measure something accurately and includes some features such as (be unbiased and free of error), consistently, and entirely (own construct and content validity), and obtain an adequate sampling of the domain). Accordingly, tests should be fair. In the present study, the theoretical background of fairness in testing and biased testing has been elaborated. Furthermore, the relationship between fairness and validity has been taken into account.

Introduction
In the recent decades, learning language as a second or foreign language is a considerable and inevitable process in an individual's life. In acquiring this process, all learners are not same and they will be encountered with different problems like understanding and applying language learning issues. To observe the level, progression, weaknesses and strengths of learners in this process, it is necessary to evaluate them. Accordingly, testing is a prominent processing language teaching and learning. According to Shohamy (2000), tests are “very powerful instruments which can determine the future of individuals and programs,” yet the “actual uses of tests may be different from the purposes they were intended for” (p. 15).

Thus, the role of testing in language cannot be ignored. According to Birjandi and Mosallanejad (2010), a test is considered as an artificial logical construct created by people. An ideal and good test would measure something accurately and includes some features such as (be unbiased and free of error), consistently, and entirely (own construct and content validity), and obtain an adequate sampling of the domain). In addition, it would always be applied for its appropriate, specified aim and would always be easily and correctly interpreted. It is easy to understand how difficult it is to achieve this, and also it is not simple in attempting to reach this ideal. An operationalized test takes mere a representative sampling of the complicated real world realms to be tested. Examinees with different language backgrounds and also with different educational backgrounds may be advantaged or disadvantaged in foreign language testing. Various performances might be based on native language distance from the target language, or might be based on unfamiliar format, or to emphasize on a skill domain that is not stressed in the educational system of the examinee's home country. Then, in the natural world, it looks that test writers and developers cannot be fair, in the ideal sense, but they can attempt to be equitable.

In the milieu of high-stakes decision-making, test fairness among stakeholders are important and they involve achievement, admission, and certification. Camilli (2007) points out test fairness has been the issue of much discussion in educational study, specifically with the recent worldwide trend towards applying testing and assessment for standard-based educational reforms (Gipps & Stobart, 2009). According to Elder (1997), a test which relates to the materials taught is said to be fair to the examinees. If, for instance, a test involves points that ought to have been covered somehow overlooked in the instruction, the test is unfair to the participants. In the Iranian context, it is assumed that most language tests in high stakes are not fair, because they do not have validity. In the present study, the theoretical framework of test fairness and test bias will be taken into account.
High-Stakes Tests

Often, to categorize, select, and judge individuals, large-scale high-stakes tests are applied as a tool. According to Bachman and Purpura (2008), these tests serve as “both door-openers and gatekeepers” (p. 456) to involve individuals in, or exclude individuals from, academic contexts and professional associations. As test takers believe high-stakes testing is considered as a door-opener to a pathway that leads to educational opportunities, resources, and success. In other words, administrators, teachers, policy makers and academic association officials tend to treat high-stakes testing as a gatekeeping means. They believe high-stakes testing is applied to specify if test takers achieve a certain proficiency and competency level.

Fairness in Language Testing

History and Definitions of Fairness

According to Gipps and Stobart (2009), in the history of language testing, fairness is considered as a prominent issue of discussion concerning with high-stakes testing. As Xi (2010) states that fairness is often associated with another fundamental concept in high-stakes testing as test validity. According to Angoff (1993), the fairness quest in educational assessment has initiated since at least the 1960s. However, language testing has widely designed on educational measurement and psychometrics for much of its methodological section, recently investigators in language testing have only begun to explicitly involve fairness issues in their discussions of the needs of a proper test development and use policy.

Different scholars have defined fairness in different ways. In this section, some of these definitions will be considered. According to Kunnan (2000), fairness as “the quality of treating people equally or in a way that is right or reasonable”. In similar words, the Merriam Webster Dictionary (2005) defines the term “fairness” as “marked by impartiality and honesty: free from self-interest, prejudice, or favoritism” (p. 342). It can be said that a test may be expressed to be fair if it is "free of bias" or "favoritism". Particularly, a test can be fair if it does not unduly advantage any groups. According to Kane (2010), indeed, this can be the traditional definition in educational measurement. It is assumed that there is a difference between fairness and justice. In the next section, the difference between fairness and justice is taken into account.

Four Possible Meanings of Fairness Based On The Standards

McNamara (1991) provided four views about fairness meaning. The first meaning is associated with requiring equal group consequences. This definition was rejected by the standards, because it notes that when the differences of the group should trigger heightened scrutiny for probable bias sources (that is a systematic error that variously affects the different groups of test takers' performance), consequence differences in and of themselves do not reveal bias. Later, it notes that there is a wide agreement which testees with equal level on the construct of interest should, on average, receive the same score regardless of group membership.

The second meaning of fairness indicates fairness as the equitable treatment of all examinees. Equal treatment based on the conditions of testing, permit to practice materials, performance feedback, retest chances, and other characteristics of test administration, involving supplying logical accommodation for test takers with disabilities when suitable equitable treatment are basic dimensions of fairness under this perspective.

The third meaning is related to require that testees have a similar chance to acquire the subject matter covered by the test. Although, the standards refer that this view is most widespread in the educational achievement testing field and that opportunity to learn usually plays no role in specifying the fairness of employee selection procedures.
The last meaning is related to the lack of predictive bias. According to this view, predictor use as fair if a common regression line can be applied to depict the predictor-criterion relationship for all subgroups of interest; subgroup variants in regression slopes or intercepts signal predictive bias.

**Fairness versus Justice**

As McNamara and Ryan (2011) point out there is a difference between fairness and justice. They believe that fairness is as being related to the technical (mostly psychometric) qualities of the test. Similar to the traditional definition of fairness alluded to earlier, they viewed fairness as "the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test-takers and the adequacy of the representation of the construct in test materials and procedures" (p. 163). In other words, justice is related to the social outcomes of test use along with the value implications and the ideologies of socio-politics forming the foundation of the test constructs. Fairness can be explained based on the Messick's (1989) validity matrix (Table 1). Fairness is associated with the evidential foundation of test use and interpretation, in the top row of Messick's matrix. In other words, justice is related to the consequential basis, including the value implications and social outcomes of test use and interpretation.

**Table 1. Messick's (1989) Validity Matrix**

<table>
<thead>
<tr>
<th>Test Interpretation</th>
<th>Test Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidential basis</td>
<td>Construct validity</td>
</tr>
<tr>
<td>Consequential basis</td>
<td>Value implications</td>
</tr>
<tr>
<td></td>
<td>Construct validity + Utility/Relevance</td>
</tr>
<tr>
<td></td>
<td>Social Consequences</td>
</tr>
</tbody>
</table>

The above mentioned dichotomy is similar to the difference which Shohamy (1997) explained. He makes between two main repertoires of bias in language tests: "those related to the test itself, such as method effects and those associated with the consequences and uses of language tests" (p. 341).

In language testing, the studies mostly focus on fairness rather than justice. Different techniques have been applied to study different dimensions of fairness. These variables can be gender, language background, academic background and content knowledge (Ryan & Bachman, 1992). On the other side, language testers have not paid attention equally to justice. The studies about justice are mostly associated with the social consequences of the test use (the fourth cell in Messick’s matrix) with less attention to the value implications (the third cell), at least as far as empirical research is concerned.

**Fairness and Validity**

According to Xi (2010), fairness has been imagined in different ways and is considered as one of the central issues in dealing with validity. In testing, fairness can be defined in different ways, but at its center is the idea of comparable validity. A fair test is comparably related to the valid inferences from individual to individual and group to group. An assessment task is taken into account biased if construct-irrelevant characteristics of the task result in different meanings for different subgroups. Fairness is associated with both validity of a given test as an index of ability and to the whole testing process as it reflects to social equity.

There are three different approaches to describe the relationship between fairness and validity across the investigated documents. As the first approach indicates that test fairness as an overarching test quality which includes diverse dimensions involving validity. JAC (1993) published *Principles for Fair Student Assessment Practices for Education in Canada* and also JCTP (2004) published *Code of Fair Testing Practices in Education*. They point out that fairness is regarded as a principal consideration in all testing dimensions and evidence in terms of technical qualities like as reliability and validity are sections of fair testing practices. It is emphasized that validity and reliability are two main considerations for good testing practice in whole contexts. These two concepts reveal that how fairness issues are integrated into each
stage of the testing process. Kunnan (2008) showed this approach in his work. Therefore, the presented fairness conceptualization indicates that test fairness is overarching and subsumes reliability and validity.

Based on the second approach, fairness is considered as a separate notion from validity. These two terms, validity and fairness, are side by side; although, how validity and fairness associate with each other remains vague. The Principles of Good Practice declares that the tenet of high-quality assessments is to give a guarantee that “assessments are valid and that all test takers are treated fairly” (University of Cambridge, ESOL Examinations, p. 8). In The ETS Standards for Fairness and Quality (ETS, 2002), fairness and validity were placed in two chapters. However, some references may be made to validity, ETS (2002) emphasizes equality and the significance of taking groups of test takers into consideration. It regards fairness as “the extent to which the test is appropriate for members of different groups, and the extent to which test takers are treated the same way, regardless of gender, race, ethnicity, and the like” (p. 71). This second approach appears to separate validity and fairness, acknowledging that fairness and validity are both basic. Nevertheless, this approach does not explain the actual distinctions between the two notions.

The third approach refers that fairness and validity to be closely interrelated (Willingham & Cole, 1997). They state that “validity evidence like as construct underrepresentation, construct irrelevant variables, test takers’ or raters’ response processes, relations of test scores to other criteria, and uses of testing “are discussed (p. 141). Then the corresponding fairness issues for each type of validity evidence are clarified and expanded in a separate chapter, with some attention given to different subgroups of test takers. This connection between discussions of fairness and validity suggests “a strong possibility for linking fairness with validity in a principled way” (Xi, 2010, p. 152). Basic issues discussed in validity are also reflected in different dimensions of fairness studies. The presented approach is agreeable to scholars like as Willingham and Cole (1997), Gipps and Stobart (2009) and Xi (2010). While acknowledging that validity traditionally concentrates more on within-group variation and fairness on between-group variation, this approach regards issues being addressed under validity and fairness to be basically the same (Camilli, 2007).

The Scope of Test Fairness
Providing a clear scope of issues for fairness investigations is one of the significant contributions of guidelines, standards and codes of practices, however the scope may be inconsistent from one document to another. Most of standards and codes of practice restrict testing activities scope or roles of the involved parties across the testing process (e.g., EALTA, 2006; JAC, 1993). With paying attention to these bounds in the testing process, ultimately the documents amalgamation specified four basic realms: test design and development, administration, scoring, and score-based test use. Apart from these common realms, two other less common areas were specified: opportunity to learn (OTL) and impact of test use. Only two documents emphasized opportunity to learn (ACACA, 1995) and four discussed the significance of positive impact of test use (2006; ETS, 2005). The realms seem to supply a comprehensive picture regarding different views on the scope of test fairness. In the following section, the testing process which involves four common areas, followed by two less common areas are discussed.

The Testing Process
The standards, guidelines, instructions and codes of practice indicate procedures to address issues of fairness across the testing process. To explain, the process of testing involves activities directly associated with testing including (1) the design of test and development, (2) administration, (3) scoring, and (4) score-based test use. It excludes what precedes a test as well as its social outcomes (Xi, 2010).

Test Design and Development
As the first realm of test design and development shows, the standards, instructions, and practice codes are generally consistent regarding the significance of practices like as choosing and training item writers, pretesting/field testing, evaluating item performance (i.e., Differential Item Functioning), conducting
fairness/sensitivity reviews, and examining comparability across parallel and modified forms. Generally, fairness in test design and development emphasizes on the examination of group differences (Xi, 2010). There is consensus among the documents that a lack of bias and comparability across different forms are basic components for supplying fairness in test design and development. The specifications of the test depict test features like as test content, format, and length should take all test takers into consideration and prevent disadvantaging specific groups of test takers. Choices of what to test among competing constructs and the formats and forms to test the content are basic dimensions to make a decision if groups of test takers will have fair chances to carry out. "By gauging narrowly defined construct(s) or introducing irrelevant variance, a test may underestimate (or overestimate) the competence of groups who would have done better (or poorer) on what was not included” (Willingham & Cole, 1999, p. 81).

**Administration**

The second area shows test administration, regular fairness practices are related to the requirement for steadfast administration procedures, test accommodations, and security. Also, these procedures must be designated to reduce cheating. There are two foci in test administration. On one hand, all the documents state that test administration, like as standard time limits, proctoring to ensure there are no irregularities, and testing physical conditions, should be as similar as probable for all test takers, taking any form of the test at any time and location. Standardized administration is desired to make sure a test is administered under constant circumstances unless there are defensible reasons to adjust the administration procedures. In other words, the majority of the documents point out the importance of special accommodations. The accommodations of the test, which suggest various treatments during test administration, are regarded to be basic to make sure test fairness (e.g., ILTA, 2007).

**Scoring**

As the third realm indicates that the issues of fairness must be regarded is scoring. In general, the standards, instruction, and practice codes agree that fairness needs consistency and impartiality in scoring for all test takers. Fair scoring hints that all test takers must be treated impartially with the same scoring procedures and schemes. Usual fairness practices in scoring involve standard scoring procedures, training of raters, score communication/reporting, and appeal systems if there are any concerns about test scores (Xi, 2010). The scoring procedure and rater’s training are significant to keep accuracy and consistency of scores of the test.

**Score-Based Test Use**

The last area was score-based test use. This area indicates that the documents widely point out the significance of fairness in the use of scores. Fairness in test score use involves considerations like as suggested interpretations, proper uses and communication, deterring misuse, stimulating different sources in decision-making, standard errors of measurement of the test, and reflecting the extent to which the test is performing its specified objectives. Only two documents supply particular information as to what constitutes fairness in decision-making: the 1999 Standards (AERA, APA, & NCME, 1999) and the Principles for Fair Student Assessment Practices for Education in Canada (JAC, 1993).

**Bias in Testing**

According to McNamara (1991), a biased test or an item is taken into account when one specific part of candidate population is advantaged or disadvantaged by some test or item characteristics which are not related to what is gauged. Bias indicates skewedness and unfair inclination toward one side (Group, population) to the determinant of other. Bias is directly associated with fairness in popular usage and assessment: A bias judgment considers factors other than those that should be informing that (McNamara, 1991).

Assessment bias can be considered as construct irrelevant variance that deform the test outcome and makes results depended on the score less valid. Systematically, bias harms one group by expanding the
score of one group and preventing the other group score. A variant index between subgroup in a test reveals bias.

Bias Sources
A biased test does not appear from a test or a test item, it originated from different sources within a testing process involving:

- administrating procedures,
- test direction,
- test content,
- test knowledge selection,
- testing method,
- rating/scoring,
- score interpretation,
- And norm sample selection (McNamara, 1991).

Bias vs. Fairness (Moral Context)
According to McNamara (1991), morality is related to a test is bias for or against a particular group of test takers that might lead to unfair decision. It should be mentioned that term bias and fairness are closely relevant but distinct at the same time. Bias is viewed as a statistical feature of the test score. Or of the prediction based upon those scores. Bias is expressed to exist when a test involves systematic sources of error in measurement or prediction. The existence of bias can be defined empirically and determined statistically. By examining the data, one can specify the extent to which a test provides bias measure or bias predictions.

On the other side, fairness is associated with a value judgment regarding decisions or actions taken as the test scores outcomes. It involves a comparison between the decision that were made and the decision that should have been made. A test is most likely to be attacked as unfair when:

- It leads to adverse decisions for some groups in the population;
- It is the sole bases for decision;
- The consequences of doing poorly on the test are harsh (Seymour, 1988).

One way to allay the unfairness result is multiple assessment through which a lot of related factors can be considered. Another way is to employ multiple phase decision models rather than making irreversible decision about everyone at the point of testing (Cronbach & Glesser, 1965).

Ethics Codes and Practice in Language Testing
Internationally, there are two basic applied codes exist in language testing: Language Testing association's Code of Ethics (2000) and draft code of practice (International Language Testing Association, 2005). The Codes of ethics and practice go beyond individual test makers and provides guidelines for language testing profession. The increase transparency of language testers works, and supply more framework for tester's work, although this framework is often ambiguous and has no built-in mechanism to deal with violation.

Code of Ethics
Codes of ethics are the moral and social aspects /consideration included in the construction and use of tests. Test constructors and users should be knowledgeable about the science of human behavior to guard against the unwarranted inferences in their interpretations of test scores. Generally speaking every profession has some ethical obligation to the public. This obligation in testing involves: Professional competency, confidentiality, privacy, objectivity, fairness. The ethics code are related to the relationship between the tester and test taker in the following tenets: 1) it emphasizes that test must not discriminate against test takers or exploit the power that they hold over them and they must obtain the test takers informed consent for any test related research. 2) Also, testers must keep the confidentiality of test takers
personal information and only depend upon those who are authorized to obtain this information. 3) The code needs language testers to update their skills and to educate others in the principle of good practice in language testing. 4) The code also requires the testers to encourage ethical and professional conduct in their profession, improve the quality of language testing and make it available as widely as possible.

Punch (1994) argues that ethical issues involve consent, deception, confidential, privacy, and equal chance to acquire. According to Shohamy (1997), the testing ethics as an extension of critical pedagogy or more carefully in testing as critical language testing. This a new notion in language testing involves a series of strategies for controlling and minimizing the misuses of a test. The proponent of critical language testing claim that large scale standard testing is the "agent of cultural, social, political and ideological agendas. According to Brown (2004), the issues (principles underlie CLT) of critical language testing are:

- Psychometric traditions are challenged by interpretive and individual procedures for predicting evaluating ability
- Testers are deeply embedded in culture and ideology
- Test takers are political subject in political context
- Test designers have responsibility to offer different modes of performance to account for varying styles and ability among test takers

Conclusion
The role of testing cannot be neglected. One side of language testing is fairness. It means that an ideal and good test would measure something accurately and includes some features such as (be unbiased and free of error), consistently, and entirely (own construct and content validity), and obtain an adequate sampling of the domain. In the present study, the main purpose was to review test fairness and biased test. Fairness is viewed as the quality of treating people equally or in a way that is right or reasonable. Fairness can be related to the validity. In this study, three approaches of relating fairness and validity were discussed. In the next section, the scope of fairness was taken into account. In the last section of this study, bias and code of ethics have been regarded.

Conflict of interest
The authors declare no conflict of interest.

Reference


