Journal of Education and Teacher Training Innovation (JETTI)
Published by Pusmedia Group Indonesia

Journal of Education and Teacher Training Innovation
Volume 3 Issue 1 Year 2025 Pages 106-125
e–ISSN : 3025-3500
DOI: 10.61227
e-mail : jetti.pusmedia@gmail.com
https://journal.pusmedia.com/index.php/JETTI

Reactions of EFL Learners to an Unfairness Test

Hossein Salarian [1*]

[1] University of Tehran, Tehran, Iran

Correspondence e-mail: Hos_Salarian@ut.ac.ir

Abstract:
The matter of bias is more muddled by the problem of obviously differentiating components of culture and instructional background from the language capacities we want to measure. This study investigates the reaction of English as a Foreign Language (EFL) learners to perceived test unfairness, examining the role of social class and gender in test performance. Drawing on fairness theories and using the reading section of the IELTS test, thirty-four intermediate-level learners were initially pretested and grouped according to gender and social class. After exposure to a uniform instructional treatment, participants completed a posttest. For this purpose, both descriptive and inferential statistics were used to compare the results of the groups. In order to make sure that the differences between the groups are not statistically significant, a one-way ANOVA was employed. Results revealed significant disparities in performance linked to social class and gender, suggesting the presence of test unfairness or differential access to the skills required. This study underscores the importance of validating assessment instruments not only in terms of construct validity but also in terms of fairness across sociocultural demographics. There are some theoretical and practical implications involved in this study.

Keywords: Bias, fairness, unfairness, ethical test preparation, social class

Article info: Submitted : 2025-06-07 | Accepted : 2025-07-08 | Published : 2025-07-09

How to Cite :

Introduction

Testing is a multi-faceted field and 'fairness' is a very complicated and potentially argumentative concept with many possible definitions in language testing. The language testing literature has gravitated to consider 'fairness' under the title of *bias* (Davies, 2010). Fair assessment actions in instructional settings impact learners' motivation, self-regulation, and more importantly instructors' believability. However, the issue has been under the influence of instructional stakeholders' various outlooks and views (Ahmadi Safa & Bahare Nasiri, 2025). In addition, Murillo and Hidalgo (2020) also focused on the instructors' conception and intended to discover how the instructors conceptualize a fair assessment.

In the same vein, Kunnan (2010) argued that test fairness as one of the most basic notions in assessment was at the center of examinations and considerations in the scope of language assessment in 1990s. He broadens the argument, referring to the AERA/APA/NCME Standards (1999) where fairness is discussed under the following heads: equitable treatment of all examinees; freedom from bias; equality of testing outcomes; and equity of opportunity to learn the testing content. And he notes that the Standards explicitly reject the idea that fairness means equality of testing outcomes for minority groups. It is individuals who may be compensated not groups. Nevertheless, test bias research is addressed to determine and where possible decrease the impact of any extraneous *and unmeasured* variables on test scores, by modifying the test' (Elder, 1997). Gipps (1990) also regarded the fairness concepts as 'equity' in which testing has conventionally been applied as an instrument of equal opportunities in instuction. As Madaus (1992c) argued that in considering the equity of alternative assessments in a high- stakes policy-driven-exam system policy should be created that construct a first and more important level functioning field for leaners and schools. Only then can we argue that a national examination system is an equitable technology for deciding about learners, or schools. The same issue is also chosen by Baker and O'Neil (1994). Equity, or fairness, is a complex area in assessment; it is not just to do with the design of tests and assessment material. The traditional psychometric approach to testing operates on the assumption that technical solutions can be found to solve problems of equity with the emphasis on using elaborate *techniques* to eliminate biased *items (*Murphy, 1990). Bachman (2010) pointed out that *bias* in association with assessment is generally involved in the assessment which is unfair to a specific group or another. Yet, this fairly simple definition determines the intricacy of the fundamental situation.

Bachman (2010) argues that *bias* in relation to assessment is generally taken to mean that the assessment is unfair to one particular group or another. This rather simple definition, however, belies the complexity of the underlying situation. The process of validation is addressed to specific test uses and specific groups of test takers. But within these groups there may be subgroups that differ in ways other than the language ability of interest. These differences may affect their test performance, and hence the validity of inferences we make on the basis of the test scores. It is important to note that differences in group performance in themselves do not necessarily indicate the presence of bias, since differences may reflect genuine differences between the groups on the ability in question. When a test happens to be in favor of some test takers due to their individual or group characteristic, we say test is biased and in cases systematic differences in test results occur that to be associated with characteristics *not* logically related to the ability in question, we should completely explore the possibility that the test is biased.

On the other hand, Mislevy (2018) asserted that test fairness considers a rational basis for adopting the interests and background knowledge of the test takers and emphasizes the significance of comprehending the manner in which a test is used based on the individual(s), its background, task, and settings. Addressing assessment practices as fair, rather than performing as instruments for classification, they should be conceived as instruments for diagnosis; rather than acting as external tools to gauge learners' practice, they should assist to develop learners' learning, and above all instead of acting as means of penalizing those who fall short of the needed standards, they should level out the overall learners' assessment (Peters et al., 2017).

Moreover, differential performance on a test, i.e. in which various groups obtain diverse score levels, cannot be the outcome of bias in the assessment; it can be because of actual differences in performance among groups which may in turn be because of various access to learning, or it can be due to actual differences in the group's achievement in the topic under investigation (Bachman, 2010). Wood (1987) explained these various (actors as the opportunity to acquire talent (access issues) and the opportunity to show talent to good effect (fairness in the assessment). Learners may blame their cheating behavior on unfair tests and/or professors. Some theories on motivational suppose that learners are more probably to cheat when they conceive the exams and tests as very unfair (e.g., needing knowledge of material that was not covered before or skills they haven't performed). Whether the unfairness is real or only perceived is not important in terms of its effect on student behavior. So, when a new situation arises, they either lack the general concept you expected they had learned or lack the skill of identifying key ideas. If a learner's knowledge organization shows a superficial comprehension of the material, problems presented in a different context might look unfamiliar and therefore unfair.

Based on the aforementioned discussion the following research questions are provided:

H1: Is there any statistically significant difference between upper social class and middle social class through unfair test?

H2: Is there any statistically significant difference between male upper social class and female upper social class through unfair test?

H3: Is there any statistically significant difference between male middle class and female middle class?

H4: Is there any statistically significant difference between male upper class and male middle class?

Review of Literature

1. The Test Fairness Framework

Test fairness as a basic issue in the assessment of tests has been in the forefront of considerations in the scope of language assessment from the late 1990s.

The Test Fairness Framework (TFF), like Rawls' (2001) view of justice as fairness- which is the view that when considering justice as fairness what is required is a veil of ignorance regarding members of society so that just and fair practices can be devised (without consideration for any group)- is framed for a well-ordered society in which there is social cooperation between citizens who are free and equal and the primary goal is social and political justice (Kunnan, 2010)

The principles and sub-principles for the TFF are as follows:

1. the Principle of Justice: a test should be fair to testees (sub-principle 1: a test should involve comparable construct validity of score interpretations and decisions; sub-principle 2: a test should not be biased based on construct-irrelevant issues); and
2. the Principle of Beneficence: a test should bring good to society (sub-principle 1: a test should develop good to society; sub-principle 2: a test should no impose harm to society).

These principles and sub-principles can then be applied in test assessment in which the particular emphasis can be on the scopes like lack of test bias, validity of test score interpretations and decisions, test access, test administration and test consequences. Moreover, test fairness considers fair testing practice in order that tests are favorable and not harmful to society (Sebolai, 2014).

In addition, the TFF that test developers and test users design a test and testing practice that is fair to all test takers without reference to a particular set of test takers. In other words, many of the accounts of fairness go back to the philosopher John Rawls and his influential argument that justice is fairness (Rawls, 2001). Rawls puts forward two principles. The first is that everybody has the same claim to the fundamental liberties. The second is that where there are inequalities they should meet two conditions, that the positions should be open to everybody in terms of equality of opportunity, and that the least-advantaged members of society should benefit most from these inequalities.

2. Aspects of Fairness

All scholars without exceptions acknowledge the continuance of unfair items. Yet, Cole and Zieky (2001) argued that there is no generally specific definition of fairness regarding testing. In fact, the definition should include a procedure for distinguishing between fair and unfair items, a procedure which is based on some measure. If a suitable procedure existed, it would supply a definition of unfair items. Some measurement

scholars apply the expression "unfair item" in narrow notion meaning that such item advocates one group of testees over another even though the groups are of comparable capacity. Differential item functioning method (DIF) verifies whether a test item is biased against particular gender, ethnic, social, or economic groups (e.g. Holland, 1988). DIF attempts to evaluate item functioning by comparison of the item scores for two or more groups. DIF cannot check item fairness on a single group; as a result it cannot offer a fairness scale.

Xi characterizes test fairness in terms of three prevailing views with examples: (1) fairness as a fairly independent test quality or general testing practice which is not specifically linked to validity (e.g., 1999 Standards); (2) fairness as a comprehensive test quality which includes various components composing validity (e.g., Kunnan, 2004, 2008), and (3) validity as the basic test quality that connects fairness directly to it (e.g., 1999 Standards, Willingham & Cole, 1997). Xi dismisses the first two views: the first view 'does not provide a mechanism for prioritizing them and for weighing one piece of fairness evidence against another' (p. xxx) and the second view: 'current validation frameworks have provided means to address all the fairness qualities proposed in Kunnan (2004) in a coherent way within the framework of a validity or assessment use argument. It does not seem necessary to treat them as separate facets of fairness' (p. xxx). These assertions help Xi to take up the third view as appropriate for her approach and illustration. Xi hitches this 'new' approach to an argument-based approach; thus coming up with a fairness argument in a validity argument.

Furthermore, there can be various facets of fairness at any one time. These facets can compose of differential item functioning (DIF) exploring sub-groups (Kunnan, 2000; Ferne & Rupp, 2007), the impact of construct-irrelevant testee characteristics on test performance (Alderson & Urquhart, 1985a, b; Zeidner, 1986; Hale, 1988; Kunnan, 1995; Clapham, 1998; Taylor et al., 1998), the impact of interviewer behavior on test-takers' speaking scores across studied groups (Brown, 2003), the impact of gender bias in oral interviews (O'Loughlin, 2002), the invariance of factor structures of test scores across groups (Swinton & Powers, 1980; Hale et al., 1989; Oltman et al., 1990; Ginther & Stevens, 1998; Stricker et al., 2005), and the reliability of multiple-choice test scores across L1 groups (Brown, 1999).

There is a range of hypotheses which are used to account for group differences (for example, between boys and girls) in performance. These are broadly either environmental or biological. Biological hypotheses for differences in group performance suggest that there are underlying psychological *differences* between groups either genetic, hormonal or due to differences in the structure of the brain. These factors are no longer used to account for differences in performance between ethnic groups, since this

approach has been discredited, but are still used to account for differences between the sexes. Environmental hypotheses include cultural, social and psychological influences (sometimes called *psychosocial variables*) that affect the development of individuals within specific groups. Both biological and environmental hypotheses imply that the differences in performance are real. Differential performance, however, may be due to factors in the test itself; in other words, groups (or individuals) may have an equal level of knowledge or skill but be unable to show it to the same extent because the test is unfair to one group. This might be because of the language used or artefacts in the illustrations or text which make the task meaningful to one group but less meaningful to another (Nisbet, 2019). As Goldstein (1993) maintained, bias is improved into the test developers' construct of the topic and their expectations of differential performance. Moreover, the improvement of conceptual frameworks for fairness in language testing has widely developed the scope of fairness (Kunnan, 2004).

3. Ethical Test Preparation

Ethical matters in language testing have been one of the crucial research subjects and problems in the scope of language testing in the 21st century (Gao & Lio, 2023). The issue, and this is the same *for* any country or setting, is what sort of test preparation is appropriate and ethical. Of course, in a high- stakes setting teachers will prepare pupils for tests and exams and the aim is that this preparation will enhance their scores. The danger is that in teaching too directly to the items and tasks on the test or questions from old exam papers, performance is enhanced narrowly, only on the test items or exam questions (Zhao et al., 2016).

What Smith (1991) tried to do was to look at test preparation from the teachers' point of view — their meanings-in-action — and from her work produced this typology:

a. no special preparation;
b. teaching test-taking skills;
c. exhortation, for example, getting a good night's sleep and doing 'your very best';
d. teaching the content known to be covered by the test;
e. teaching to the test, i.e. using materials with the same format and content;
f. stress inoculation — this refers to boosting pupils' confidence and feelings of self-efficacy so they can do well on the tests and feel good about themselves '. . . work on test preparation primarily to inoculate their pupils against emotional paralysis in the face of the tests and against feelings of stupidity that the tests seem to engender';
g. practicing on items of the test itself or parallel forms;
h. cheating — the teachers had very precise views about what constituted cheating: giving extra time on tests; providing hints on - rephrasing questions; providing

correct answers; altering marks on answer sheets; providing hints on rephrasing questions.

Cannell (of the Lake Wobegon report) considered all but the first two of them as cheating (unethical) since they would lead to inflated scores. To take this point of view, Smith (1991) pointed out that one should believe that the standardized test (not taught to) accurately shows the construct (and that teaching to the test results in inflated scores which cannot be generalized to the broader construct). Wiggins (1989) considered the argument for authentic assessment into the equity field by arguing that the one-off standardized test that treats all testees as the same, is inherently inequitable. Wiggins also claims that unreliability—the reason usually given for not using teachers' assessments of pupils — is only a problem when teachers operate in private without shared criteria. Therefore, teaching to the test in any form is viewed as 'unethical' when it only develops scores and not performance on the basic construct; And, when only some instructors or schools conduct this way in an accountability setting.

4. Validity and Fairness

Validation involves an evaluation of the credibility, or plausibility, of the proposed interpretations and uses of test scores (Cronbach, 1971; Messick, 1989; Kane, 2006). Effective validation, therefore, depends on a clear, explicit statement of the proposed interpretations and uses, with the statement including a specification of the population and of the range of contexts in which the interpretations and uses will occur.

In addition, validity and fairness are closely connected. An assessment that is unfair, in the sense that it systematically misrepresents the standing of some individuals or some groups of individuals on the construct being measured or that tends make inappropriate decisions for individuals or groups is, to that extent not valid for that interpretation or use. Similarly, an assessment that is not valid in the sense that it tends to generate misleading conclusions or inappropriate decisions for some individuals or groups will also be unfair (Nisbet, 2019).

Likewise, validity theory has tended to focus on the accuracy and appropriateness of score-based interpretations and decisions about all of the individuals in the population of interest. Analyses of fairness have tended to focus on group differences and on differences in the accuracy and appropriateness of interpretations and decisions across groups, which are defined in terms of race/ethnicity, gender, age, and so on. The issues being addressed are basically the same (Mazzoli Smith et al., 2018).

In the same yay, performance assessment is a particular type of assessment, though not of course restricted to language. Lam suggests that the search for fairness in

performance assessment can pit equality against equity. The demand for equality is that all students should be assessed in a standardized manner using identical assessment methods and content and the same administration, scoring and interpretation procedures. This makes sense but even the most trivial difference between groups can lead to the claim of bias (Lowenberg, 1998). The demand on the other hand for equity is that every student should be assessed as an individual with regard to instruction context and personal history (prior knowledge, cultural experience, language proficiency and so on). But individualizing, however equitable it may appear, makes it difficult to ensure comparability of results. Lam's conclusion appears to be that validity rather than fairness is the prize: 'How much better off are we with assessments that are equally invalid for all groups (fair but invalid) than assessments that are invalid for some groups (valid but unfair)?' (Lam, 1995, p. 4).

What's more, fairness is a key issue in gender studies and since gender is universal, the issue of gender fairness is an exemplar of all such issues elsewhere. Fairness, writes Childs (1990), 'refers to the ways test results are used' and she points out that 'test publishers go to great lengths to make sure … that the recommended uses of the tests are not likely to be unfair to members of one gender'. For Childs, gender bias in testing is the result of construct irrelevant factors: 'the result of characteristics of the examinees that are stable (such as gender or race) and that are characteristics other than those the test is intended to measure' (1990, p. 1).

A potential pitfall in any bias detection procedure is in the choice of suitable control measures of ability. The validity of the criterion used as benchmark for comparing different groups of test-takers is difficult to ascertain and this means that the results of both test-external and test-internal bias detection must be interpreted with caution. External bias-detection methods usually use the procedure of regression analyses of scores of one test against those achieved on another measure considered to be obtaining the same ability. The difficulty with test-internal bias analyses (i.e., procedures for identifying differential item functioning [DIF]) is that they are circular (Camilli, 1993).

Due to the probably biased nature of the standard applied to investigate test bias, it is usually favored that any discrepancies coming from a statistical analysis be subjected to expert scrutiny to make sure the source of these discrepancies (Herman & Cook, 2019). As researchers such as Spurling and Ilyin (1985) and Shepard (1982; 1987) have pointed out, bias detection requires the exercise of *judgement* to determine whether observable differences in group performance are the result of measurement error or rather of *real* differences in the ability under test. Scrutiny of this kind sometimes reveals that discrepancies in item functioning are *not* in fact due to construct-irrelevant variance.

Method

Participants

The selected participants of this study comprised learners with intermediate proficiency level in a language institute in Tehran, Iran. These thirty-four learners were selected from among 60 learners through administering a Placement Test as pre-test. Those who scored 31 in Grammar & Vocabulary, 8 in Reading, and 8 in Writing, based on Oxford Placement Test and its criteria, were considered intermediate. Then, using match-paired design, these students were divided into four equal groups, 15 each. It should be mentioned that these learners were native speakers of Farsi and studied English as a foreign language. Their age ranged from 23 to 31 with the average age of 27.

Instruments

At first and before the treatment, the Oxford Placement Test (OPT) was used in order to identify the learners language skills and sub-skills to ensure about the homogeneity of the participants and to select intermediate level learners. The second instrument used in a pretest and posttest was the reading section of IELTS with acceptable reliability (i.e., 0.74) and validity as the material, a multiple-choice test with 40 items testing situations. The validity of this instrument was considered by some English experts who were three university teachers with more than 10 years of experience**.**

Data Collection Procedure

This study followed a true experimental design. The participants were selected randomly and experienced similar, but not identical conditions, except that participants in the experimental group receive the treatment. On day one, the reading section of an IELTS Test was administered to the learners as pre-test, and according to the test results, learners who got the criteria and passed the test were selected for the purpose of the study in order to get homogenous learners based on their abilities. After that, they were divided into five groups based on their sexes (male/ female), and social classes (lower, middle, upper). Because only one of them was from the lower social class, he was omitted from the rest of the study and the participants were divided into four groups. After treatment, posttest was administered to these four groups.

Data analysis

The collected data was used to see if the four groups involved are benefited the same from the situations or not. For this purpose, both descriptive and inferential statistics were used to compare the results of the groups. For descriptive statistics, the mean, standard deviation, as well as minimum and maximum scores were provided. As for the inferential statistics, since there were four groups involved in the study, one-way ANOVA and the related post hoc tests were employed.

Results

Descriptive Statistics for Oxford Placement Test (OPT)

In order to place and homogenize the learners in groups in terms of their proficiency and based on the rubric of the OPT for the score bands, an Oxford Placement Test developed by Allen (2004) was administrated. Table 1 shows the number of learners, mean and standard deviation.

Table 1.
Results of Descriptive Statistics of OPT

Group	N	Mean	Std. Deviation	Std. Error Mean
Experimental	30	45.23	12.044	1.13
Control	30	46.75	13.376	1.36

The learners with scores 1 SD above and 1 SD below the mean were selected from the population. As displayed in the Table 1, the two groups were to some extent at similar levels of language proficiency since their means are rather similar. For determining the significance of the difference between the mean values of these two groups, an independent-samples t-test was used. Table 2 provides the results of this test:

Table 2.
Independent-Samples t-test of the Performances of the Experimental Group and the Control Group on the OPT

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	2.468	.102	.879	89	.376	1.346	1.072	-1.202	3.241
Equal variances not assumed			.879	47.15	.238	1.346	1.072	-1.193	3.41

As seen in Table 2, the p-value (.102) in the result of the *Levene's Test for Equality of Variances* was greater than .05. Thus, the first line of results (i.e. Equal variances assumed) was used. The p-value .376 (displayed as Sig. 2-tailed) in this line of results was greater than .05, too. As a result, there can't be viewed a significant difference between the performances of the groups on the OPT and they were homogeneous in terms of their language proficiency before the onset of the treatment. Also, the Shapiro-Wilk test was used as the numerical tool for assessing the assumption of normality.

The Results of the Pretest

In order to make sure that the participants were homogeneous with regard to their reading ability; the researcher pre-tested them. After that, their scores were compared to see if they were homogeneous. Their means are very close to each other. However, in order to make sure that the differences between the groups are not statistically significant, a one-way ANOVA was employed. Table 3 shows the results of this ANOVA.

Table 3
The Results of the One-way ANOVA on Reading Pretest

Source	SS	df	MS	F	Sig.
Between Groups	.138	3	.046	.117	.950
Within Groups	29.750	66	.391		
Total	29.888	69			

According to Table 3, the amount of F-observed (.117) is not statistically significant (p= .950). Therefore, it can be said that the four groups' reading performances on the pretest were similar and they can be considered homogeneous.

Results of the Posttest

After the treatment, the IELTS learners post-tested. Then their reading abilities were scored and the results were compared with each other. Table 4 presents the descriptive statistics for the posttest.

Table 4.
Descriptive Statistics for the reading Posttest

Group	N	Mean	SD	Min	Max
Upper class, (male)	15	7.15	.933	5	7
Upper class, (female)	15	6.30	.865	3	6
middle class, (male)	15	8.30	.923	4.5	6.5
middle class, (female)	15	7.10	.788	2.75	6
Total	**60**	**7.21**	**1.122**		

By checking the information in Table 4, one can see that there are differences between the four groups. In order to find out whether or not these differences are statistically meaningful, a one-way ANOVA was run. Table 5 indicates the results of this ANOVA.

Table 5.
The Results of the One-way ANOVA on Reading Posttest

	SS	df	MS	F	Sig.
Between Groups	40.638	3	13.546	17.523	.000
Within Groups	58.750	66	.773		
Total	99.388	69			

Table 5 asserts that the amount of F-observed (17.523) is significant at the probability level of p= .000 which is indicative of a statistically significant amount. However, this table does not tell where the exact place(s) of difference(s)

is/are. To find this out, a Scheffe post hoc test was employed. Table 6 reports the results of this post hoc test.

Table 6.
The Results of the Scheffe Post hoc Test for the Posttest

Groups	Mean Difference	Sig.
Upper Class	.85*	.031
Middle Class	.05	.998
Upper Class, (Male)	1.20*	.031
Upper Class, (Female)	.75*	.000
Middle Class, (Male)	-1.15*	.001
Middle Class, (Female)	1.20*	.001
Upper Class (Male)	-.80	.998
Middler Class (Male)	.-1.20*	.048

The data in Table 6 reveals the following facts about the differences between the four groups:

1. There is a significant difference between Upper class, and middle class (p= .031, mean difference= .85). This result rejects the first null hypothesis which states that, "Upper class and middle class have been influenced the same effect by unfair test", and since the mean difference is positive, it can be said that Upper class had a better performance than middle class.
2. There is a significant difference between male Upper class and female Upper class (p= .001, mean difference= 1.20). This result rejects the second null hypothesis which states that, "male Upper class and female Upper class have been influenced the same effect by unfair test", and since the mean difference is positive, it can be said that male Upper class outperformed implicit group.
3. There is a significant difference between male middle class and female middle class (p= .001, mean difference= -1.15). This result rejects the third null hypothesis which states that, "there is no significant difference between male middle class and female middle class", and since the mean difference is negative, it is evident that explicit-virtual group benefited better from the feedback than explicit-physical group.
4. There is a marginally significant difference between male Upper class and female middle class (p= .048, mean difference= -.80). This result rejects the forth null hypothesis which states that, "there is no significant difference between male

Upper class and male middle class", and here again, since the mean difference is negative, it is evident that male Upper class performed better than male middle class.

Discussion

The first research question investigated whether there is a statistically significant difference between upper and middle social class learners in their performance on an unfair test. The analysis revealed that learners from the upper class performed significantly better than those from the middle class. This supports the hypothesis that socioeconomic background can influence language test outcomes, especially when the test context reflects experiences more familiar to higher social classes. These findings are consistent with the notion that unequal access to quality educational resources can manifest in testing environments. When the test reflects knowledge or cultural experiences that align more with the lifestyles of upper-class learners, it inherently creates an advantage for them. Consequently, this raises ethical concerns about the content and context of standardized tests, especially in multicultural or economically diverse societies.

This outcome supports the Test Fairness Framework (TFF), specifically the sub-principle that a test should not be biased due to construct-irrelevant factors. In this case, the influence of social class appears to be a construct-irrelevant variable that unfairly benefits one group over another. The significant performance gap thus challenges the validity of the test in accurately reflecting reading ability alone. Moreover, this result underlines the importance of equitable test development and administration. Ensuring that the materials are culturally and socially neutral is critical to maintaining fairness. If students from lower or middle classes face unfamiliar themes, vocabulary, or contexts in reading passages, their performance will likely suffer—not due to language ability but because of irrelevant external factors. Therefore, the answer to the first research question is affirmative: there is a statistically significant difference between upper and middle social class learners' test performance, pointing to the presence of socioeconomic bias in the assessment tool used in this study.

The second research question focused on the potential performance differences between male and female learners within the upper social class group. The results revealed a statistically significant difference, with males outperforming females. This suggests that gender, even within the same social class, can play a critical role in test outcomes, possibly due to varying educational experiences, cultural expectations, or test anxiety levels. From a theoretical standpoint, this finding echoes research emphasizing that test fairness must account for gender as a relevant variable. It also supports the idea

that gender-based societal roles might influence familiarity with certain test content, or confidence in a testing environment. Male learners might have had more exposure to test-related scenarios or less social pressure impacting their performance.

The observed disparity could also reflect differential classroom treatment or instructional methods that favor male learning styles or behavioral expectations. Teachers may unconsciously provide more support or higher expectations for one gender, which accumulates over time into measurable performance differences on standardized assessments.

Furthermore, this gap poses a challenge to the interpretation of test scores as fair and valid indicators of language ability. If female learners consistently perform lower due to social conditioning or systemic bias in education, test scores fail to reflect their true proficiency. This leads to misinformed decisions based on the assessment results, violating the principles of justice and beneficence outlined in the TFF.

Hence, the significant gender-based difference observed in the upper class group confirms that fairness in assessment must consider both social class and gender. Addressing such disparities requires not only more inclusive test design but also broader educational reforms aimed at minimizing gender-based inequities.

The third research question explored whether male and female learners in the middle social class differ significantly in their test performance. The data indicated that male learners outperformed their female counterparts, pointing once again to gender-based discrepancies in testing outcomes, despite both groups belonging to the same socioeconomic tier.

This finding aligns with broader concerns in educational literature regarding the systemic challenges female learners face, especially in high-stakes testing environments. Factors such as gender stereotypes, classroom dynamics, or societal expectations may all contribute to the differences observed. Test anxiety or lack of self-efficacy, often reported more frequently among female learners, may have also played a role.

The implication here is that fairness issues extend beyond socioeconomic status and into the realm of gender equity. If test design or classroom instruction unintentionally benefits one gender over another, then the resulting scores are not valid reflections of learners' abilities. This undermines both the construct validity and ethical legitimacy of the test outcomes.

The results suggest the need for greater scrutiny of how test items are framed, ensuring they are not implicitly favoring one gender. Furthermore, assessment conditions should be structured to support equitable performance, possibly including measures such as balanced representation in examples, inclusive topics, and test accommodations where necessary.

In conclusion, the significant performance gap between male and female middle class learners emphasizes that even within a single socioeconomic stratum, gender can influence test fairness. It is essential to investigate and address these layered sources of bias to ensure just and valid assessment practices.

The fourth research question assessed whether there is a statistically significant difference between male learners from upper and middle social classes. The findings revealed that male upper-class learners outperformed their middle-class peers, suggesting that even within the same gender, social class continues to impact performance outcomes on an unfair test.

This reinforces the argument that socioeconomic factors contribute to differential access to learning resources and experiences that are crucial for test success. For example, upper-class learners may benefit from enriched educational environments, additional tutoring, or culturally familiar test content—all of which enhance their readiness for assessments like IELTS.

This result further illustrates that fairness cannot be presumed based on gender or group homogeneity alone. It must be contextually examined across multiple dimensions, including social class. In this case, although gender is constant, the variance in class reveals the continued influence of background variables unrelated to language ability.

From a fairness standpoint, such differences violate the principle of construct relevance. The fact that social class, an unrelated factor to linguistic skill, can sway results highlights the urgent need to reevaluate test constructs and content. Test developers must ensure that the tasks presented are neutral and accessible to all learners, regardless of background.

Therefore, the evidence supports the rejection of the null hypothesis and confirms that male learners from upper and middle classes experience the effects of social bias differently. This finding substantiates the broader claim that social equity in assessment must be actively engineered rather than assumed.

Conclusion and Implications

This study was an attempt to find out the reaction of EFL immediate learners to an unfairness test. Test bias is a sophisticated issue. It may take a broad range of shapes, composing of the misinterpretation of test scores, unequal prediction of criterion performance, sexist or racist content, unfair content regarding the experience of testees, on suitable selection procedures, inaccurate criterion measures, and threatening settings and situations of testing. In language tests, the scope of bias is more complex through the problem of specific distinctive components of culture and educational background from the language capacities we wish to measure. Measurement professionals cannot come to

an agreement on the definition of the term "item fairness". To any amount the unfairness measure deviates from zero, to that amount the item is unfair. When the measure raises the cutoff score, the item is determined as specifically unfair. Childs (1990) argued that fairness 'refers to the manners test results are used' and she asserted that 'test providers try to make sure that the recommended applications of the tests are not likely to be unfair to members of one gender.

As it was illuminated in the preceding section of the study, regarding the results of the posttest the findings of the study revealed that first, upper social class and middle social class have different effects on EFL students' immediate learning of reading section of IELTS, and since the mean difference is positive, it can be said that high class had a better performance than middle class. Also, male upper social class outperformed female middle social class. Furthermore, it was evident that male middle social class benefited better from this test than female middle social class. And finally, it was proved that male Upper class performed better than male middle class.There are theoretical and practical implications for this study. Nevertheless, because of some limitations of this study, its applications and efficient findings may not be generalizable to all conditions. Therefore, further studies are required to improve it.

References

Ahmadi Safa, M. & Nasiri, B. (2025).Fairness in language classroom assessment practices: what do EFL teachers underscore?. *Language Testing in Asia, 15*(1).49-62. Article number: 1 (2025)

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

Bachman, L. F. (2005). Building a test use argument. *Language Assessment Quarterly, 2,* 1–34.

Backman, L. F. (2010). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Boyd Jenkins, O. (2003). The meaning of "fairness". *The Virtual Research Centre: Links for Learning*. http://www.orvillejenkins.com/faithlife/fairnessfl.html

Childs, R. A. (1990). Gender bias and fairness. *ERIC Digest, ED328610.* http://www.ericdigests,org/pre-9218/gender.htm

Camilli, G. (2006). *Test fairness*. In R. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 221–256), Westport, CT: American Council on Education and Praeger.

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (2008). (Eds.), *Building a validity argument for the Test of English as a Foreign Language.* Mahwah, NJ: Lawrence Erlbaum**.**

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

Herman, J., & Cook, L. (2019). *Fairness in classroom assessment*. In S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement (pp. 243–264). Routledge.

Holcombe, R. G. (1983). Applied fairness theory: comment. *The American Economic Review, 73*(5), 1153–1156.

Gao, S. & Liu, J. (2023). Ethical issues in language testing from a validity perspective. *Open Journal of Social Sciences, 11*(4), 143-158.

Gipps (1990), *beyond testing*. Harvard University Press.

Jones, K., Evans, C., Byrd, R., Campbell, K. (2000) Gender equity training and teaching behavior. *Journal of Instructional Psychology, 27*(3), 173-178.

Kane, M. T. (2006). *Validation*. In Brennan, R. L. (Ed.), Educational measurement, 4th ed.(pp. 18–64). Washington, DC: American Council on Education/Praeger.

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta et al. (Ed.), C*urrent developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: 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, UK: Cambridge University Press.

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, UK: Cambridge University Press.

Kunnan, A. J. (2010). Introduction: Test fairness, test bias and DIF. *Language Assessment Quarterly, 4*(2), 109–112.

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

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

Lam, T. C. M. (1995). *Fairness in performance assessment*. ERIC Digest, ED391982. http://www.ericdigests.org/1996–4/fairness.htm

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

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mazzoli Smith, L., Todd, L., & Laing, K. (2018). Students' views on fairness in education: The importance of relational justice and stakes fairness. *Research Papers in Education, 33*, 336–353.

Murillo, F. J., & Hidalgo, N. (2020). Fair student assessment: A phenomenographic study on teachers' conceptions. *Studies in Educational Evaluation, 65*, 1–10.

Nisbet, I. (2019). Fairness takes center stage. *Assessment in Education Principles Policy and Practice, 26*(1), 111–117.

Peters, R., Kruse, J., Buckmiller, T., & Townsley, M. (2017). It's just not fair! Making sense of secondary students' resistance to a standards-based grading. *American Secondary Education, 45*(3), 9–27.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13.

Porter, T. (2003). Measurement, objectivity, and trust. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 241–255.

Rawls, J. (2001). *Justice as Fairness: A restatement*. Cambridge, MA: Harvard University Press.

Sadker, D., Sadker, M. (1994). *Failing at Fairness: How Our Schools Cheat Girls*. Toronto, ON: Simon & Schuster Inc.

Saunders, A. (2008). *Egalitarianism and fairness*. *The Philosopher's Zone*. ABS Radio. June 2008. http://www.abc.net.au/rn/philosopherszone/stories/2008/2275460.htm

Shohamy, E. (2000). Fairness in language testing. In Kunnan, A. J. (Ed.), *Fairness and validation in language assessment* (pp. 15–19). Cambridge, UK: Cambridge University Press.

Velasquez, M., Andre, C., Shanks, T., & Meyer, M. J. (2008*). Justice and fairness*. Markkula Center for Applied Ethics. Santa Clara University, California. http://www.scu.edu/ethics/practicing/decision/ justics.html

Willingham, W. W. & N. Cole (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

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

Zhao, M. R., Mu, B. L., & Lu, C. P. (2016). Teaching to the test: approaches to teaching in senior secondary schools in the context of curriculum reform in China. *Creative Education, 7*, 32- 43.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing, 3*, 80–95.

Zeidner, M. (1987). A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*, 55–71.